Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Joint Probabilistic Modelling of Images and Non-imaging Covariates: A Causal Perspective

Daniel COELHO DE CASTRO

Supervisor
Dr Ben GLOCKER

Submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Computing of Imperial College London

September 2020

Declaration of Originality

I, Daniel Coelho de Castro, hereby declare that this thesis is my own work, unless otherwise specified with appropriate citations, and that all published material reproduced herein is in accordance with the relevant copyright licenses.

Copyright Statement

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution 4.0 International Licence (CC BY).

Under this licence, you may copy and redistribute the material in any medium or format for both commercial and non-commercial purposes. You may also create and distribute modified versions of the work. This on the condition that you credit the author.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Abstract

Typical machine-learning tasks in imaging applications involve training a model to predict some target annotation (e.g. a class label) from an image. However, by neglecting valuable side-information that is often available in real-world settings, such models are vulnerable to confounding and may not generalise outside the lab environment. This thesis argues for the integration of causality into image analysis workflows, supported by the development of methodologies for jointly modelling images and non-imaging data, especially for healthcare applications. We first demonstrate that causal reasoning can shed new light on some of the greatest challenges in medical image analysis: data scarcity, i.e. the limited availability of high-quality annotated data; and data mismatch, whereby exogenous domain differences or sample selection may threaten the external validity of a predictive model. Turning to practical aspects of probabilistic modelling, we verify the suitability of nonparametric mixtures to modelling image intensities, and employ them in a novel intensity normalisation method. We then leverage similar techniques to develop a unified probabilistic model for a (benign) face recognition application, incorporating face features, unlimited identities, sparse and noisy name labels, and contextual side-information. To facilitate more general research into learning with heterogeneous data, we additionally introduce an experimental sandbox based on images of handwritten digits, including measurable shape attributes and controllable image perturbations. Finally, bridging causality theory and deep probabilistic modelling, we propose the framework of deep structural causal models, which offer tractable generative, interventional, and counterfactual capabilities for high-dimensional, heterogeneous data. Along with the increased robustness of models that can naturally integrate multiple sources of data, further considerations surrounding transparency, interpretability, explainability, and fairness make it clear that a causal perspective contributes towards building machine-learning systems that are not just highly performant, but also safe and reliable.

Contents

1 Introduction **13**
- 1.1 Motivation . . . 13
- 1.2 Research aims . . . 15
- 1.3 Thesis outline, contributions, and publications . . . 15

2 Background **19**
- 2.1 Imaging . . . 19
- 2.2 Probabilistic modelling . . . 21
  - 2.2.1 Bayesian nonparametrics . . . 22
  - 2.2.2 Variational inference . . . 29
  - 2.2.3 Flow-based models . . . 33
- 2.3 Causality . . . 37
  - 2.3.1 The causal hierarchy . . . 37
  - 2.3.2 Causal reasoning . . . 39
  - 2.3.3 Causal graphical models . . . 41
  - 2.3.4 Building a causal diagram . . . 42

3 Causal reasoning for medical image analysis **45**
- 3.1 Introduction . . . 45
- 3.2 Causality matters . . . 47
  - 3.2.1 Predictive analytics in medical imaging . . . 47
  - 3.2.2 Causality in medical imaging . . . 48
  - 3.2.3 Worked clinical examples . . . 49
- 3.3 Data scarcity . . . 50
  - 3.3.1 Tackling data scarcity via semi-supervision . . . 50
  - 3.3.2 Tackling data scarcity via data augmentation . . . 52
- 3.4 Data mismatch . . . 52
  - 3.4.1 Data mismatch due to dataset shift . . . 53
  - 3.4.2 Data mismatch due to sample selection bias . . . 56
- 3.5 Discussion . . . 57

4 Modelling and normalising image intensities **63**
- 4.1 Unsupervised segmentation . . . 64
- 4.2 Nonparametric intensity model . . . 65
  - 4.2.1 Intensity model . . . 65
  - 4.2.2 Inference . . . 65
- 4.3 Segmentation experiments . . . 67
  - 4.3.1 Evaluation . . . 67
  - 4.3.2 Normal tissue segmentation . . . 67

4.3.3 Lesion discovery . . . 68
4.3.4 Model visualisation . . . 69
4.4 Intensity normalisation . . . 71
4.5 Nonparametric density matching . . . 71
4.5.1 Intensity model . . . 72
4.5.2 Density matching . . . 72
4.5.3 Warping . . . 74
4.5.4 Practical considerations . . . 76
4.6 Normalisation experiments . . . 77
4.6.1 Dataset . . . 77
4.6.2 Setup . . . 77
4.6.3 Histogram fitness . . . 79
4.6.4 Tissue statistics . . . 79
4.6.5 Centre classification . . . 81
4.7 Conclusion . . . 81

5 A probabilistic model for contextual face recognition 83
5.1 Introduction . . . 84
5.2 Related work . . . 86
5.3 A model of identities . . . 87
5.3.1 Context model . . . 87
5.3.2 Identity model . . . 88
5.3.3 Face model . . . 89
5.4 Semi-supervised label model . . . 90
5.4.1 Label prior . . . 91
5.4.2 Label likelihood . . . 92
5.4.3 Label prediction . . . 93
5.5 Evaluation . . . 93
5.5.1 Experiment 1: Unknown person detection . . . 93
5.5.2 Experiment 2: Identity discovery . . . 95
5.5.3 Experiment 3: Semi-supervised labelling . . . 97
5.6 Extensions to the context model . . . 98
5.6.1 Unbounded contexts . . . 98
5.6.2 Environmental cues . . . 99
5.7 Conclusion . . . 100

6 Morpho-MNIST: A sandbox for learning with heterogeneous data 103
6.1 Introduction . . . 103
6.1.1 Contributions . . . 105
6.1.2 Related work: Datasets . . . 106
6.1.3 Related work: Quantitative evaluation . . . 107
6.2 Morphometry . . . 108
6.2.1 Processing pipeline . . . 109
6.2.2 Stroke length . . . 110
6.2.3 Stroke thickness . . . 110
6.2.4 Slant . . . 111
6.2.5 Width and height . . . 112
6.3 Perturbations . . . 112
6.3.1 Global: Thinning and thickening . . . 112
6.3.2 Local: Swelling . . . 113
6.3.3 Local: Fractures . . . 113
6.4 Evaluation case studies . . . 113

6.4.1 Sample diversity . . . 114
6.4.2 Disentanglement . . . 116
6.4.3 Predictive model evaluation . . . 121
6.5 Discussion and perspectives . . . 124
6.6 Conclusion . . . 125

7 Structural causal models for images and non-imaging covariates 127
7.1 Introduction . . . 128
7.2 Deep structural causal models . . . 129
7.2.1 Structural causal models . . . 129
7.2.2 Deep mechanisms . . . 130
7.2.3 Deep counterfactual inference . . . 132
7.2.4 Discrete counterfactuals . . . 134
7.3 Related work . . . 135
7.4 Case study 1: Morpho-MNIST . . . 136
7.4.1 Data generation . . . 136
7.4.2 Experimental setup . . . 137
7.4.3 Results . . . 140
7.5 Case study 2: Brain imaging . . . 144
7.5.1 Data generation . . . 144
7.5.2 Experimental setup . . . 145
7.5.3 Results . . . 146
7.6 Conclusion . . . 150

8 Conclusion 151
8.1 Summary of achievements . . . 151
8.2 Limitations, open questions, and outlook . . . 153
8.2.1 Validity of causal models . . . 153
8.2.2 Beyond fully observed, fully specified causal models . . . 154
8.2.3 Causal understanding of high-dimensional features . . . 155
8.2.4 Practical implications and applications . . . 156

Bibliography 157

A Face recognition supplement 177
A.1 Label likelihood . . . 177
A.2 Gibbs sampler conditionals . . . 178

B Morpho-MNIST supplement 183
B.1 Morphometrics of plain and perturbed datasets . . . 183
B.2 Perturbation examples . . . 186

List of Figures

2.1 Orthogonal slices of an example brain MRI scan . . . 20
2.2 Posterior predictive distributions for two realisations of a DP with base distribution $H = \mathcal{N}(0, 1)$ . . . 25
2.3 Behaviour of the number of DP clusters as the number of samples increases . . . 26
2.4 Illustration of the stick-breaking process . . . 27
2.5 Behaviour of stick-breaking processes with different concentrations . . . 27
2.6 Plate diagram for a Dirichlet process mixture model . . . 29
2.7 Plate diagrams for a general class of probabilistic models suited to variational inference 30
2.8 Simple causal diagrams illustrating fundamental multivariate relationships . . . 39

3.1 Key challenges in machine learning for medical imaging . . . 47
3.2 Causal diagrams for medical imaging examples . . . 49
3.3 Selection diagrams for causal and anticausal dataset shift scenarios . . . 54
3.4 Causal diagrams for different sample selection scenarios . . . 56
3.5 A 'scaffold' causal diagram summarising typical medical imaging workflows . . . 58

4.1 Results for the normal tissue segmentation experiment . . . 68
4.2 Uni- and multimodal grey matter (GM) segmentations of IXI subject 52 . . . 68
4.3 Lesion segmentation results and example . . . 69
4.4 Empirical and fitted five-dimensional distributions for TBI subject 17770 . . . 70
4.5 Comparison of two MRI scans, before and after the proposed NDFLOW normalisation . 72
4.6 Example density flow between two histograms . . . 77
4.7 Intensity distributions, colour-coded by imaging centre . . . 78
4.8 Histograms and Q–Q plots of each of the methods against the target histogram . . . 78
4.9 Centre prediction statistics . . . 80

5.1 Conventional and proposed face recognition paradigms . . . 85
5.2 Overview of the proposed generative model, encompassing the context model, identity model, face model, and label model . . . 87
5.3 Hard label predictions of the proposed semi-supervised label model and nearest-neighbour classification . . . 91
5.4 Results of the unknown person detection experiment on test images . . . 94
5.5 The simulation used in Experiment 2, showing identities coming in and out of the camera frame . . . 96
5.6 Identity clustering consistency . . . 96
5.7 Label prediction accuracy . . . 97
5.8 Overview of the extended probabilistic model . . . 98

6.1 Overview of Morpho-MNIST . . . 105

6.2 Stages of the image processing pipeline . . . 109
6.3 Distribution of morphological attributes in the plain MNIST training dataset . . . 110
6.4 Measuring shape attributes . . . 111
6.5 Distribution of morphometric attributes for MNIST test dataset and samples from some generative models . . . 115
6.6 Partial correlations between inferred latent codes and morphometrics of test images . . . 118
6.7 Partial correlations between 1000 sampled latent codes and morphometrics of the corresponding generated images . . . 119
6.8 InfoGAN latent space traversals . . . 120
6.9 Partial correlations between inferred latent codes and morphometrics of test images for InfoGAN-C . . . 120
6.10 Results for toy outlier detection experiments showcasing data stratification by stroke thickness . . . 123
6.11 Application of the proposed morphometrics to the USPS dataset . . . 124

7.1 Classes of deep causal mechanisms considered in this work . . . 131
7.2 Random exemplars from the synthetically generated Morpho-MNIST test dataset . . . 137
7.3 Computational graphs of the structural causal models for the Morpho-MNIST example 138
7.4 Comparison of the target covariates and the corresponding values measured from the generated images . . . 140
7.5 Random samples generated by the independent, conditional and full models . . . 141
7.6 Conditional samples generated by the independent, conditional, and full models . . . 141
7.7 Reconstructions . . . 141
7.8 Distributions of thickness and intensity in the true data, and learned by the full and conditional models . . . 142
7.9 Difference between conditioning and intervening, based on the trained full model . . . 142
7.10 Counterfactuals generated by the full model . . . 143
7.11 Additional counterfactual examples from the full model . . . 143
7.12 Random examplars from the test set of the adopted UK Biobank dataset . . . 144
7.13 Brain imaging example . . . 145
7.14 Densities for the true data (KDE) and for the learned model . . . 147
7.15 Random samples from the model trained on the UK Biobank dataset . . . 148
7.16 Conditional samples from the model trained on the UK Biobank dataset . . . 148
7.17 Original samples and reconstructions from the model trained on the UK Biobank dataset 148
7.18 Original samples and counterfactuals from the model trained on the UK Biobank dataset 149

B.1 Distribution of morphological attributes for plain MNIST digits . . . 183
B.2 Distribution of morphological attributes for thinned MNIST digits . . . 184
B.3 Distribution of morphological attributes for thickened MNIST digits . . . 184
B.4 Distribution of morphological attributes for swollen MNIST digits . . . 185
B.5 Distribution of morphological attributes for fractured MNIST digits . . . 185
B.6 Examples of globally thinned digits . . . 186
B.7 Examples of globally thickened digits . . . 186
B.8 Examples of digits with local swelling . . . 187
B.9 Examples of digits with local fractures . . . 187

List of Tables

3.1 Types of dataset shift 54
3.2 Types of sample selection 56
3.3 Step-by-step recommendations 59

4.1 True posterior and variational distributions for the multivariate DPGMM 66
4.2 Tissue statistics after normalisation 79
4.3 Centre classification accuracy 80

6.1 Kernel two-sample tests between model samples and true test data 116
6.2 Settings for InfoGAN disentanglement experiments 118
6.3 Example results on digit recognition (accuracy, %) under synthetic domain shift using the proposed data perturbations 122
6.4 Example results on some of the new MNIST prediction tasks 123

7.1 Comparison of the associative abilities of the models shown in Fig. 7.3 140

Acknowledgements

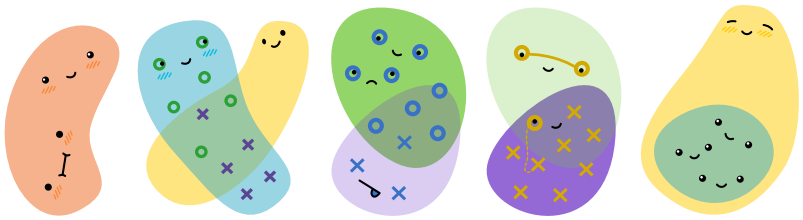
I would first like to thank my parents, Ludmila and Newton, and sister, Helena, back home in Rio for their unconditional love and support—even after I (once again) escaped 10 000 km away over a certain big puddle—and for quite literally giving me the conditions to make it this far. A special word of appreciation also goes to my would-be-were-it-not-for-the-apocalypse wife, Lucy, for creating a home with me in this foreign country and for keeping me fed and loved when things get rough (and for bringing bunnies into my life).

I would like to express my huge appreciation to my supervisor, Ben Glocker, who has been an outstanding mentor and role model throughout these past five years. Thank you for your kind, unwavering support and invaluable guidance—since your leap-of-faith in offering me this PhD opportunity back in 2014, up to our continued collaboration moving forward.

I am hugely thankful to Sebastian Nowozin for scouting me and providing a tremendously enriching internship experience at Microsoft. Also a big thank-you to Tobias Rijken for giving me the fruitful opportunity to apply my research ideas in a real-world context at Kheiron.

Special gratitude also goes to my close collaborators Nick Pawlowski, Miguel Monteiro, Kostas Kamnitsas, Qi Dou, and Ian Walker, with whom I was able to explore new research directions and make a handful of contributions to the field. Lastly, a huge collective thank-you to all my other friends and colleagues from the HiPEDS CDT and BioMedIA lab, who helped make this whole experience honestly a lot of fun!

My doctoral work was partly supported by the CAPES Foundation, Ministry of Education, Brazil (Finance Code 001, Full Doctorate Abroad Scholarship BEX 1500/2015-05), and also received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 757173, project MIRA, ERC-2017-STG). All UK Biobank data used in this work is accessed under Application Number 12579.



My sister's reimagining of Fig. 3.1. © 2020 Helena Coelho de Castro

Chapter 1

Introduction

1.1 Motivation

Imaging is an incredibly powerful data modality, providing dense, spatially structured measurements in diverse application domains, e.g. consumer photography, remote sensing, microscopy, robotics, and medical imaging. Notably, the latter has since its inception become a vital healthcare tool for diagnosis, screening, monitoring, and treatment planning.[1] When one has access to large collections of images, machine learning may be employed for predictive tasks by automatically extracting patterns that are associated with a target annotation of interest (e.g. class labels or object contours). However, by neglecting valuable complementary side-information that is often available in real-world settings, such models are vulnerable to confounding and may not generalise outside the lab environment.

As an example, consider an image-based medical screening task, whose goal is to detect whether a medical scan (e.g. a chest radiograph) presents an abnormality that warrants clinical investigation. A typical approach to build a machine-learning screening system would collect a retrospective dataset of images and recall decisions (i.e. whether the individual was called back to the clinic for further examination) and train a supervised discriminative model such as a convolutional neural network to predict label from image.

Why is this simplistic strategy arguably misguided? In clinical practice, a screening or diagnostic decision does not happen in a 'vacuum'—a radiologist does not just analyse the acquired imaging data on its own. Rather, they take into account the patient's full health record (including demographics, clinical data, illness and treatment history, family history, etc.), and draw on their medical knowledge to establish the findings in light of this information. This means that the model has only a very narrow

[1] Much of the research presented in this thesis is motivated by and aimed at medical image analysis applications, but the general considerations also apply to other imaging settings, *mutatis mutandis*.

view of the problem, as it is missing key information in order to replicate the true annotation process. For example, consider that certain anatomical features may appear benign in a given age range but malignant in another. Without age inputs, a model would have difficulty disambiguating such cases. In other words, the image-to-label mapping can be causally *confounded* by several important covariates (e.g. age, sex, health status) that the predictive model cannot use.[2] As a consequence, the algorithm needs to rely purely on signals present in the image for its predictions. Because the model is trained to maximise predictive power (i.e. 'How often do images that look like this have this label?') rather than to mimic reality (i.e. 'What would an expert decide for a patient like this?'), we speculate this may ultimately lead to it memorising spurious correlations.

A further problem arises when the development dataset and deployment environment have different data distributions, which we will refer to as *data mismatch*.[3] In this situation, the predictive model is likely to overfit to the characteristics of the training dataset (especially if it has high capacity, e.g. a large neural network) and perform much more poorly at test-time. As we will discuss in detail in Section 3.4, this can be caused either by differences in various exogenous factors (e.g. demographics, disease prevalence, image acquisition conditions) or by data-dependent sample selection.

In imaging, the prevalent strategies to mitigate the resulting biases generally involve some image pre-processing (like spatial alignment/registration, intensity normalisation) and/or are based on domain adaptation/generalisation (DA/DG).[4] Unlike the direct pre-processing approaches—which address only variations in image acquisition parameters—typical DA and DG methods make no explicit assumptions about the nature of the differences between domains. Instead, they presume that high-capacity black-box models can be made to generalise well with end-to-end learning algorithms. When no side-information is available beside the images and perhaps domain indicator labels, this 'brute-force' approach may be a sensible choice. However, the invariances learned in this fashion depend on the specific inductive biases encoded in the chosen model architectures and loss functions (e.g. Dou et al., 2019). On the other hand, acknowledging the causal structure of the processes behind domain discrepancies is known to lead to generalisable predictors in various applications (e.g. Bareinboim and Pearl, 2016; Subbaswamy et al., 2019), and the same can likely be achieved in imaging problems (e.g. preliminary results on the related issue of adversarial robustness are presented by C. Zhang et al., 2020).

Finally, a crucial aspect of systems intended for deployment in safety-critical applications such as

[2] Here, confounding means that the input (image) and output (label) share common causes, which introduces additional statistical dependencies unrelated to the direct relationship of interest. More details are explained in Section 2.3.2.

[3] Though this generic phenomenon is often called 'domain shift' in the computer-vision literature, we will avoid using this term as it ambiguous and could cause confusion with the specific types of mismatch discussed later (Section 3.4; see also Storkey, 2009).

[4] Domain adaptation (DA) is the setting wherein not only do we have access to data from the source domain(s), but some data from the target domain is also available (labelled or not). In contrast, domain generalisation (DG) is the case without any target-domain data, although multiple source domains may be available for training (Dou et al., 2019).

healthcare is *uncertainty* (Ghassemi et al., 2019). Real data is almost always noisy, and the correct prediction for a datum may be ambiguous even to an expert—especially when involving high-dimensional data such as images. In order for humans to trust the predictions of such an algorithm, it must generally be able to give a measure of confidence in its predictions, which may also be used to defer a consequential decision to a human. This motivates the pursuit of probabilistic models, as opposed to purely deterministic predictors.

1.2 Research aims

Given the considerations above, my doctoral research has had two complementary goals:

(A) **To establish the utility of causality in imaging problems.** In particular, I have aimed to understand how a causal perspective can illuminate issues surrounding the external validity of image-based predictive models, and how those conclusions can inform the construction of generalisable machine-learning systems and lead to more transparent reporting.

(B) **To develop effective methodologies for jointly modelling images and non-imaging data.** In order to accurately and robustly replicate complex phenomena involving imaging, we seek unified probabilistic models that can integrate multiple sources of data of different types, while also conforming to the assumed structural relationships between them. This direction of research supports the first aim by enabling the translation of causal assumptions into practical machine-learning model design.

1.3 Thesis outline, contributions, and publications

The remainder of this thesis is organised as below. At the start of each chapter, there is a box recalling the publication(s) whose content is reproduced or adapted therein.

Chapter 2 introduces the foundational concepts that motivate and underpin the contributions of this thesis. It highlights the specific challenges of handling imaging data, followed by a presentation of the probabilistic-modelling building blocks used in our work (viz. nonparametric Bayesian mixtures, variational inference, and normalising flows), and an introduction to causality theory.

Chapter 3, based on **Castro** et al. (2020), proposes a fresh perspective on some of the main obstacles to the generalisation of predictive models in medical imaging applications. Leveraging causal reasoning, we not only analyse the two prevailing techniques for circumventing scarcity of annotated data—semi-supervision and data augmentation—but also lay down a categorisation of various types of dataset mismatch according to the causal structure of the data-generating process.

- **Daniel C. Castro**, Ian Walker and Ben Glocker (2020). 'Causality matters in medical imaging'. *Nature Communications* 11, 3673. DOI: 10.1038/s41467-020-17478-w.

Chapter 4 presents two applications of Bayesian nonparametrics to intensity modelling in medical images. The first part of this chapter briefly recapitulates the work from my Master's thesis (**Castro**, 2016), later presented at a workshop (**Castro** and Glocker, 2016), which employed nonparametric mixtures for unsupervised medical image segmentation. The second part investigates another application of these models and some concepts from normalising flows to the task of intensity normalisation (**Castro** and Glocker, 2018), addressing a common type of acquisition shift (as defined in Chapter 3).

- **Daniel C. Castro** (2016). 'Dirichlet process mixture models applied to brain image segmentation'. MRes thesis. London: Imperial College London;
- **Daniel C. Castro** and Ben Glocker (2016). 'Dirichlet process mixture models: application to brain image segmentation'. In: *NIPS 2016 Workshop on Practical Bayesian Nonparametrics* (Barcelona, Spain). HDL: 10044/1/65183;
- **Daniel C. Castro** and Ben Glocker (2018). 'Nonparametric density flows for MRI intensity normalisation'. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018* (Granada, Spain). Vol. 11070. LNCS. Springer, Cham, pp. 206–214. DOI: 10.1007/978-3-030-00928-1_24. arXiv: 1806.02613.

Chapter 5 describes a probabilistic model for facial identity recognition that also integrates contextual cues. First published in **Castro** and Nowozin (2018b) and then patented (Nowozin et al., 2019), this system again uses nonparametric Bayesian components in a hierarchical model that decouples the task of face identification into identity recognition (i.e. acknowledging a familiar face) and name tagging (i.e. 'putting names to faces'). The proposed recognition module explicitly reasons about unknown individuals and can accommodate unlimited identities encountered in different contexts. The naming portion of the model works with sparse and noisy labels and is able to handle named acquaintances, unnamed familiar faces, and complete strangers. In the final part of this chapter, we discuss an extension of this model that additionally integrates contextual covariates (e.g. spatial location, time), enabling unsupervised discovery of an unbounded number of contexts (**Castro** and Nowozin, 2018a).

- **Daniel C. Castro** and Sebastian Nowozin (2018b). 'From face recognition to models of identity: a Bayesian approach to learning about unknown identities from unsupervised data'. In: *Computer Vision – ECCV 2018* (Munich, Germany). Vol. 11206. LNCS. Springer, Cham, pp. 764–780. DOI: 10.1007/978-3-030-01216-8_46. arXiv: 1807.07872;
- Sebastian Nowozin, Tom Ellis, Cecily Morrison and **Daniel C. Castro** (2019). 'Image Processing for Person Recognition'. U.S. patent 10621416. Microsoft Technology Licensing, LLC;
- **Daniel C. Castro** and Sebastian Nowozin (2018a). 'Contextual face recognition with a nested-hierarchical nonparametric identity model'. In: *NeurIPS 2018 Workshop on All of Bayesian Nonparametrics* (Montréal, Canada). arXiv: 1811.07753.

Chapter 6 introduces Morpho-MNIST, a flexible toolbox for quantitative evaluation in the context of heterogeneous data (**Castro** et al., 2019). Building on the MNIST dataset of handwritten digit images, it proposes a methodology for measuring digit shape attributes (stroke length, thickness, slant, width, and height), which can be used for empirical analysis or as additional sources of data. In addition, the framework includes a collection of morphological perturbations (thinning, thickening, swelling, and fractures) that introduce controllable factors of variation into the image space. We show with case studies that Morpho-MNIST provides novel insights for a variety of tasks spanning generative modelling, representation learning, and evaluation of predictive models.

- **Daniel C. Castro**, Jeremy Tan, Bernhard Kainz, Ender Konukoglu and Ben Glocker (2019). 'Morpho-MNIST: quantitative assessment and diagnostics for representation learning'. *Journal of Machine Learning Research* 20, 178.

Chapter 7 brings together causality theory and modelling of heterogeneous data, introducing deep structural causal models (DSCMs) as a tractable approach to counterfactual inference in high dimensions (Pawlowski*, **Castro*** et al., 2020). Our extensive experiments leveraging Morpho-MNIST (Chapter 6) demonstrate that DSCMs possess capabilities at all three levels of the causal hierarchy (associational, interventional, and counterfactual; Section 2.3.1). A further case study with brain MRI images and scalar covariates (sex, age, brain and ventricle volumes) shows that the proposed algorithm can generate surprisingly realistic counterfactual medical images that preserve low-level image details.

- Nick Pawlowski*, **Daniel C. Castro*** and Ben Glocker (2020). 'Deep structural causal models for tractable counterfactual inference'. In: *Advances in Neural Information Processing Systems (NeurIPS 2020)*. Forthcoming, *equal contribution. arXiv: 2006.06485.

Chapter 8 finally synthesises the main contributions and findings of this thesis, placing them in the context of the original aims. We moreover lay out some of the remaining open questions and limitations of this research, with pointers to future research directions.

Other works and collaborations

During the course of my PhD research, I was also involved in several collaboration projects that relate to the topics and aims of this thesis. For example, Glocker et al. (2019) studied the effects of acquisition shift (Section 3.4.1) in medical image analysis, showing that detectable scanner biases persist even after widely adopted data harmonisation pipelines. Dou et al. (2019) and Robinson et al. (2020) proposed different approaches for addressing data mismatch (Section 3.4) in imaging applications at the level of model architectures and learning algorithms. For semi-supervised learning (Section 3.3.1), Kamnitsas et al. (2018) introduced a novel regularisation algorithm based on feature

similarity graphs. Le Folgoc et al. (2019) explored a special representation of shape meshes, a useful data modality that can be extracted from and analysed alongside images to derive insights about e.g. anatomical variation. On the topic of probabilistic modelling, Schlemper et al. (2018a) employed Bayesian deep learning techniques to estimate uncertainty in MRI reconstruction (Section 2.1), and Monteiro et al. (2020) proposed a simple yet effective way to model spatially correlated distributions over image segmentation maps. A complete list is given below, in reverse chronological order:

- Miguel Monteiro, Loïc Le Folgoc, **Daniel C. Castro**, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk and Ben Glocker (2020). 'Stochastic segmentation networks: modelling spatially correlated aleatoric uncertainty'. In: *Advances in Neural Information Processing Systems (NeurIPS 2020)*. Forthcoming. arXiv: 2006.06015;
- Robert Robinson, Qi Dou, **Daniel C. Castro**, Konstantinos Kamnitsas, Marius de Groot, Ronald M. Summers, Daniel Rueckert and Ben Glocker (2020). 'Image-level harmonization of multi-site data using image-and-spatial transformer networks'. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020*. Forthcoming. arXiv: 2006.16741;
- Qi Dou, **Daniel C. Castro**, Konstantinos Kamnitsas and Ben Glocker (2019). 'Domain generalization via model-agnostic learning of semantic features'. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (Vancouver, Canada), pp. 6447–6458;
- Loïc Le Folgoc, **Daniel C. Castro**, Jeremy Tan, Bishesh Khanal, Konstantinos Kamnitsas, Ian Walker, Amir Alansary and Ben Glocker (2019). 'Controlling meshes via curvature: spin transformations for pose-invariant shape processing'. In: *Information Processing in Medical Imaging* (Hong Kong). Vol. 11492. LNCS. Springer, Cham, pp. 221–234. DOI: 10.1007/978-3-030-20351-1_17. arXiv: 1903.02429;
- Ben Glocker, **Daniel C. Castro**, Robert Robinson and Ender Konukoglu (2019). 'Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects'. In: *Medical Imaging Meets NeurIPS Workshop* (Vancouver, Canada). arXiv: 1910.04597;
- Chengzhi Shi, Ben Glocker and **Daniel C. Castro** (2019). 'PVAE: learning disentangled representations with intrinsic dimension via approximated L_0 regularization'. In: *NeurIPS 2019 Disentanglement Challenge*. eprint: https://openreview.net/forum?id=HJg8stY2oB;
- Konstantinos Kamnitsas, **Daniel C. Castro**, Loïc Le Folgoc, Ian Walker, Ryutaro Tanno, Daniel Rueckert, Ben Glocker, Antonio Criminisi and Aditya Nori (2018). 'Semi-supervised learning via compact latent space clustering'. In: *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)* (Stockholm, Sweden). Vol. 80. PMLR. PMLR, pp. 2464–2473;
- Jo Schlemper, **Daniel C. Castro**, Wenjia Bai, Chen Qin, Ozan Oktay, Jinming Duan, Anthony N. Price, Joseph V. Hajnal and Daniel Rueckert (2018a). 'Bayesian deep learning for accelerated MR image reconstruction'. In: *International Workshop on Machine Learning for Medical Image Reconstruction* (Granada, Spain). Vol. 11074. LNCS. Springer, Cham, pp. 64–71. DOI: 10.1007/978-3-030-00129-2_8;
- Jo Schlemper, Ozan Oktay, Wenjia Bai, **Daniel C. Castro**, Jinming Duan, Chen Qin, Joseph V. Hajnal and Daniel Rueckert (2018b). 'Cardiac MR segmentation from undersampled k-space using deep latent representation learning'. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018* (Granada, Spain). Vol. 11070. LNCS. Springer, Cham, pp. 259–267. DOI: 10.1007/978-3-030-00928-1_30.

Chapter 2

Background

Section 2.2.1 is adapted from Chapter 2 of my Master's thesis:

- **Daniel C. Castro** (2016). 'Dirichlet process mixture models applied to brain image segmentation'. MRes thesis. London: Imperial College London.

Section 2.3 contains material from

- **Daniel C. Castro**, Ian Walker and Ben Glocker (2020). 'Causality matters in medical imaging'. *Nature Communications* 11, 3673. DOI: 10.1038/s41467-020-17478-w.

2.1 Imaging

An image is a spatial signal representing a snapshot of one or more physical properties of a scene or object. For example, colour (RGB) photography captures photons in the 'red', 'green', and 'blue' wavelength bands that are reflected toward the detector; a radiograph represents the amount of high-energy radiation (X-rays) transmitted through an object; magnetic resonance imaging measures various properties pertaining to nuclear relaxation times in response to magnetic stimulation. The independent signals captured at each spatial location are referred to as channels, and can be understood as parallel scalar images that can be jointly analysed as multivariate images or rendered as a composite picture (e.g. a colour figure).

Not all imaging signals are acquired directly in a spatial layout, however, as the physics of some modalities require additional processing. For example, computed tomography (CT) gathers the raw X-ray signals projected from a rotating source onto many different planes, in a 'sinogram' representation (so called because it is a superposition of sinusoid curves). Using different principles, a magnetic resonance scanner measures the raw radio signal in 'k-space', which encodes its magnitude and phase

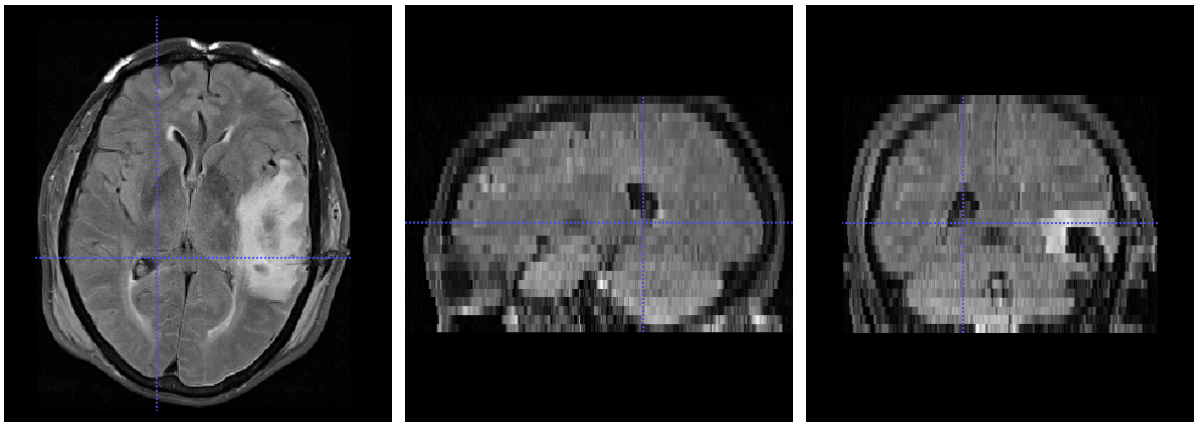


Figure 2.1 Orthogonal slices of an example brain MRI scan. This image has a strongly anisotropic resolution of $0.7 \times 0.7 \times 5\,\mathrm{mm}^3$.

across 2D or 3D spatial frequencies. The spatial images in these modalities are then reconstructed by applying (approximations of) the inverse Radon and inverse Fourier transforms, respectively.

Whereas traditionally acquired via analogue means (e.g. film), imaging signals are now predominantly captured with digital sensors—which is evidently required for any kind of quantitative computational analysis. This means not only that the originally continuous medium must be discretised, i.e. sampled at some fixed spatial resolution, but also that the corresponding continuous signals must be quantised into an integral or floating-point digital format.

Once acquired, images are represented digitally most often as numeric arrays (known as a *raster*), where values are arranged in a rectangular lattice, typically with fixed uniform spacing in each direction. Each spatially-indexed element in such an array is referred to as a *pixel* ('picture element', in two dimensions, but often used more generally) or *voxel* ('volume element', in three dimensions), and can be scalar-valued or 'number-of-channels'–dimensional (or even matrix-valued, e.g. in diffusion tensor imaging). For computational modelling, images must be viewed as mathematical variables living in a space of width × height (× depth) × number-of-channels dimensions—which may easily be in the order of millions for medical scans. Here, the intensity of each pixel in each channel corresponds to the value of a specific dimension.

Especially in medical image analysis—wherein we are often interested in quantitative measurements such as volumes (e.g. ml, cm^3) or lengths/distances (e.g. mm)—it is crucial to acknowledge the physical spacing of pixels. This is usually provided in the image file header as a tuple of lengths in millimetres along each axis and is often in the range of $1 \times 1 \times 1\,\mathrm{mm}^3$, though in some applications the resolution may be anisotropic (e.g. 'thick slices', with high 'in-plane' resolution but wider spacing in the 'off-plane' direction; see example in Fig. 2.1).

A further piece of important meta-information is the orientation of the coordinate axes, given by an origin vector and a rotation matrix that map pixel indices to fixed physical coordinates in the

scanner. For example, the 'world' coordinates $\mathbf{r}$ of a voxel indexed by (i, j, k) are given by

$$\mathbf{r}(i,j,k) = \underbrace{\begin{bmatrix} | & | & | \\ \mathbf{v}_X & \mathbf{v}_Y & \mathbf{v}_Z \\ | & | & | \end{bmatrix}}_{\text{orientation}} \cdot \underbrace{\begin{bmatrix} s_X & & \\ & s_Y & \\ & & s_Z \end{bmatrix}}_{\text{pixel spacing (mm)}} \cdot \begin{bmatrix} i \\ j \\ k \end{bmatrix} + \underbrace{\mathbf{t}}_{\text{origin}} = \underbrace{\begin{bmatrix} | & | & | & | \\ s_X\mathbf{v}_X & s_Y\mathbf{v}_Y & s_Z\mathbf{v}_Z & \mathbf{t} \\ | & | & | & | \end{bmatrix}}_{\text{'image-to-world' affine transform}} \cdot \begin{bmatrix} i \\ j \\ k \\ 1 \end{bmatrix}, \tag{2.1}$$

where $\mathbf{v}_X$, $\mathbf{v}_Y$, and $\mathbf{v}_Z$ are orthogonal unit vectors. These coordinate axes are typically aligned with the left–right, posterior–anterior, and inferior–superior directions, respectively, with one of several existing sign conventions.

These considerations highlight that images are complex objects whose processing may require special care. This is especially true when we wish to analyse a collection of medical images that are sourced from multiple acquisition sites and/or scanners. For example, the images may have different resolutions and orientations, requiring spatial alignment ('registration') and resampling to a common resolution. In addition, even if they belong to the same modality and are acquired with a similar protocol, the resulting intensity scales are unlikely to be mutually calibrated (i.e. any given intensity value corresponding to the same tissue type in both images).[5] We take a closer look at these kinds of disparities in Chapter 3, where they are discussed under the term *acquisition shift*. Chapter 4 further presents one proposed solution to harmonise differences specifically in intensity scales. Moreover, as argued in Section 1.1, modelling uncertainty is a crucial factor in certain image analysis applications. Given the complexity and high dimensionality of imaging data, its probabilistic modelling requires special considerations, which we discuss in Section 2.2 below.

2.2 Probabilistic modelling

In this section, we will review a few building blocks and methodologies that enable building expressive probabilistic models. Section 2.2.1 will give an overview of a powerful class of nonparametric models that are able to approximate arbitrary distributions and whose complexity can adapt as more data is observed, in addition to being well suited for unbounded clustering problems. We employ these versatile models in Chapter 4 to estimate image intensity distributions and in Chapter 5 to represent infinite numbers of identities, names, and contexts for face recognition. Section 2.2.2 covers variational inference, a general machinery for performing approximate learning in intractable probabilistic models. It will discuss one variety suitable for conjugate graphical models, such as the nonparametric ones in Chapter 4, and another that facilitates the incorporation of deep neural networks, which we

[5] A rare exception is CT, whose scale in Hounsfield units is standardised to known physical properties of air and water.

exploit for high-dimensional counterfactual inference in Chapter 7. Finally, Section 2.2.3 presents the fundaments of normalising flows, a flexible paradigm for defining explicit likelihood models that can be fitted via an exact training objective. For example, we propose in Chapter 7 to use these as conveniently invertible causal mechanisms.

2.2.1 Bayesian nonparametrics

Building a probabilistic model of an observed system generally means specifying a compact mathematical representation with (often interpretable) free parameters, whose values we attempt to infer from a collection of data. Specifically, consider mixture models. Given an empirical data distribution, a classic statistical mixture attempts to represent it by a finite, weighted superposition of simple distributions, called 'components'. By assigning each data point to its most likely component, such models naturally lend themselves to clustering applications—for example, they can be useful for medical image segmentation when applied to pixel intensities, with clusters roughly corresponding to tissue classes (Ashburner and Friston, 2005). One important limitation of these models, however, is the need to pre-specify the size of the parameter space (here, the number of mixture components), which requires either explicit domain knowledge or a potentially expensive model selection procedure.

How can we go beyond the typical approach, which prescribes models with some fixed number of parameters, toward defining arbitrarily flexible models? One possible solution lies in Bayesian nonparametrics. The 'nonparametric' paradigm refers not to models with no parameters, but rather to ones with infinitely many parameters! Despite their unbounded parameter space, nonparametric Bayesian models can be manipulated with computationally convenient finite-dimensional representations, when conditioned on finite amounts of data. This allows them not only to adjust their complexity to a given dataset, but also to flexibly scale up as additional data is observed. While there exist elegant nonparametric Bayesian models for learning with other infinite-dimensional objects such as functions, sequences, matrices, trees, and more (see overview in Ghahramani, 2013), in this thesis (Chapters 4 and 5) we focus on unbounded clustering structures based on Dirichlet processes, which we describe below.

2.2.1.1 Dirichlet processes

Consider a probability measure G over some measurable sample space Θ and a finite measurable partition $\{A_1, \ldots, A_K\}$ of Θ.[6] This partition implicitly defines a multinomial variable representing to which part A_k a given sample $\theta \sim G$ belongs, recalling that $p(\theta \in A_k \,|\, G) = G(A_k)$. Now let us assume

[6] A finite measurable partition is a finite collection of non-empty disjoint subsets that covers the entire set.

that G is not known, but rather is a random distribution. Because G is random, the probabilities of θ falling into each cell of the partition form themselves a random vector, $(G(A_1), \ldots, G(A_K))$.

Definition 2.1. Let α be a positive scalar and H a probability measure over Θ. If, for every finite measurable partition $\{A_1, \ldots, A_K\}$ of Θ, the vector of marginal probabilities has a Dirichlet distribution,

$$(G(A_1), \ldots, G(A_K)) \sim \mathrm{Dir}(\alpha H(A_1), \ldots, \alpha H(A_K)), \tag{2.2}$$

then we say that G follows a *Dirichlet process* with concentration α and base measure H, denoted $G \sim \mathrm{DP}(\alpha, H)$ (Ferguson, 1973).[7]

A few properties stem directly from the Dirichlet characterisation in Eq. (2.2). For any arbitrary subset $A \subset \Theta$ (for example, an element of a partition $\{A_1, \ldots, A_K\}$ as mentioned above), the mean and variance of G are given by

$$\mathbb{E}[G(A)] = H(A) \qquad \text{and} \qquad \mathbb{V}\mathrm{ar}[G(A)] = \frac{H(A)(1 - H(A))}{\alpha + 1}.$$

Intuitively, these indicate that H is the *mean* of the Dirichlet process, while α controls its variance: the higher its value, the more *concentrated* the distribution of G will be around H.

Now, assume we have some fixed partition $\{A_1, \ldots, A_K\}$. Given a set of N draws from G, $\boldsymbol{\theta}_{1:N}$, the conjugacy between the multinomial and Dirichlet distributions implies that the posterior becomes

$$(G(A_1), \ldots, G(A_K)) \,|\, \boldsymbol{\theta}_{1:N} \sim \mathrm{Dir}(\alpha H(A_1) + N_1, \ldots, \alpha H(A_K) + N_K), \tag{2.3}$$

where $N_k = |\{n : \theta_n \in A_k\}|$, the number of samples in each subset A_k.

Since the above would hold for every possible partition of Θ, we conclude that G must contain a collection of point masses.[8] In addition, we see that we have obtained the same Dirichlet form as Eq. (2.2). Therefore, the posterior distribution of G is again a DP,

$$G \,|\, \boldsymbol{\theta}_{1:N} \sim \mathrm{DP}\left(\alpha + N,\ \frac{\alpha}{\alpha + N} H + \frac{N}{\alpha + N} \frac{1}{N} \sum_{n=1}^{N} \delta_{\theta_n}\right), \tag{2.4}$$

whose posterior base distribution is a weighted average between the original H and the empirical distribution, with updated concentration parameter $\alpha + N$ (Teh, 2010).

The posterior predictive distribution is obtained by integrating out the random distribution G,

[7] Compare with Gaussian processes, whose finite-dimensional marginals are Gaussian.

[8] Recall that a point mass, atom, or Dirac measure is defined as $\delta_\theta(A) = \mathbf{1}[\theta \in A] = \begin{cases} 1, & \theta \in A \\ 0, & \theta \notin A \end{cases}$.

resulting simply in its posterior base distribution (Eq. (2.4)):

$$\theta_{N+1} \,|\, \boldsymbol{\theta}_{1:N} \sim \frac{1}{\alpha + N}\left(\alpha H + \sum_{n=1}^{N} \delta_{\theta_n}\right). \tag{2.5}$$

This means that, with probability $\frac{\alpha}{\alpha+N}$, θ_{N+1} will take a new value drawn from H or, with probability $\frac{1}{\alpha+N}$ each, it will take the same value as one of the previously drawn samples. The sequence of posterior predictive distributions for each consecutive sample is a modified Pólya urn scheme, known as the Blackwell–MacQueen urn (Blackwell and MacQueen, 1973).

The Pólya urn is a useful abstraction for understanding the behaviour of certain combinatorial stochastic processes, including the Dirichlet process. Suppose an urn contains a number of balls of different colours. At each step, we draw a ball from the urn uniformly at random and then replace it in the urn along with a new ball of the same colour.

In the DP case, we further imagine the urn initially contains α special black balls (α here assumed integral for illustration). A black ball, when drawn, prompts us to return it in addition to a ball of a new random colour. Linking back to the mathematical definitions, the 'colour' of each ball is the value of θ_n and the act of choosing a new 'colour' is equivalent to drawing a new $\theta_{N+1} \sim H$.

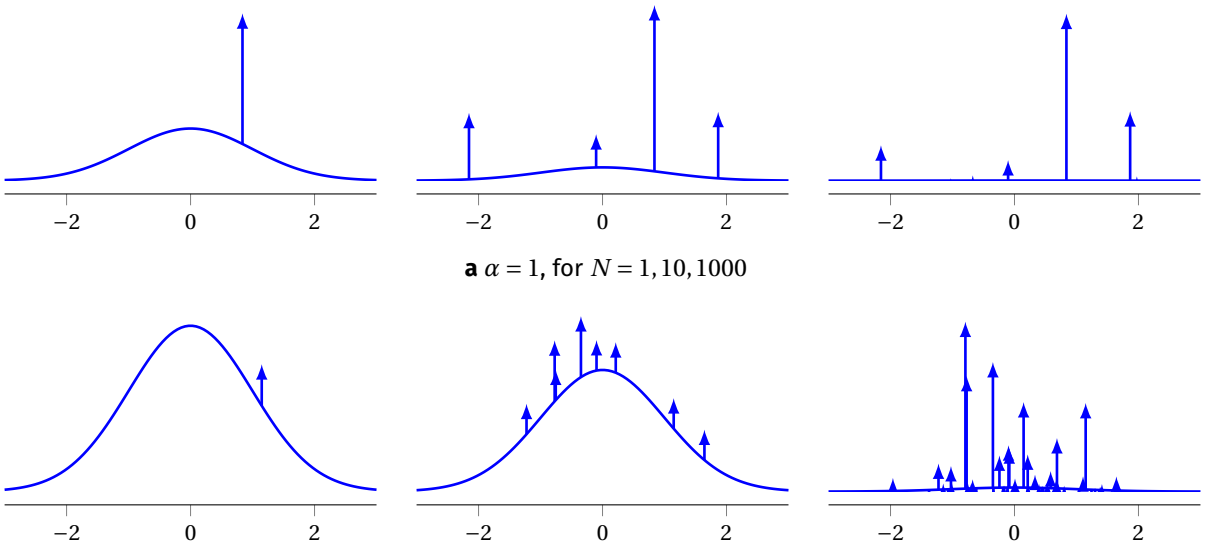
As is clear from the urn analogy, some of the samples ('balls') will have repeated values ('colours') given a long enough sequence of draws. Let $\{\theta_k^*\}_{k=1}^K$ denote the set of distinct values among $\{\theta_n\}_{n=1}^N$ and N_k the numbers of repetitions of each θ_k^*. The posterior predictive distribution in Eq. (2.5) can then be rewritten as

$$\theta_{N+1} \,|\, \boldsymbol{\theta}_{1:N} \sim \frac{1}{\alpha + N}\left(\alpha H + \sum_{k=1}^{K} N_k \delta_{\theta_k^*}\right). \tag{2.6}$$

The above expression makes evident the clustering or rich-gets-richer property of DPs: the more times a particular value θ_k^* is repeated, the more likely it is to be repeated again by future samples (namely, with probability $\frac{N_k}{\alpha+N}$).

Figure 2.2 illustrates the evolution of the posterior predictive distribution. Note how the $\delta_{\theta_k^*}$'s dominate over H as N grows large. Additionally, we can see that having a lower value of α results in much fewer distinct atoms for a given N.

Chinese restaurant process The clustering property also suggests an underlying partition of the N indices into $K \leq N$ blocks. Let us introduce a set of indicator variables, $z_1, \ldots, z_N \in \{1, \ldots, K\}$, which tell what is the unique value of each θ_n, that is, $\theta_n = \theta_{z_n}^*$. We will later refer to these as cluster assignments. The distribution of this partition over integers is known as the Chinese restaurant process (CRP;



a $\alpha = 1$, for $N = 1, 10, 1000$

b $\alpha = 10$, for $N = 1, 10, 1000$

Figure 2.2 Posterior predictive distributions for two realisations of a DP with base distribution $H = \mathcal{N}(0, 1)$**.** The arrows represent point masses.

Aldous, 1985; Ewens, 1972), written as

$$p(z_{N+1} = k \,|\, \mathbf{Z}_{1:N}) = \begin{cases} \frac{N_k}{\alpha+N}, & N_k > 0 \\ \frac{\alpha}{\alpha+N}, & \text{new } k \end{cases}, \tag{2.7}$$

where, $N_k = |\{n : z_n = k\}|$. We will denote this distribution as $\text{CRP}(\alpha, \mathbf{Z}_{1:N})$.

The analogy is explained as follows: consider a Chinese restaurant with an infinite number of tables, indexed by k. The tables are initially empty, and, when the first customer arrives, they sit at the first table. The customers are sociable and, as they arrive in sequence, each has the choice to either join one of the occupied tables, with probability proportional to the number of people already there (N_k), or to sit alone at the next empty table, with probability proportional to α (Aldous, 1985). Figure 2.3a illustrates the outcome of the CRP across a number of simulations.

We can use the CRP to construct a DP by sampling an infinite number of θ_k^*'s from the prior base distribution H and then using the CRP to determine the partitioning:

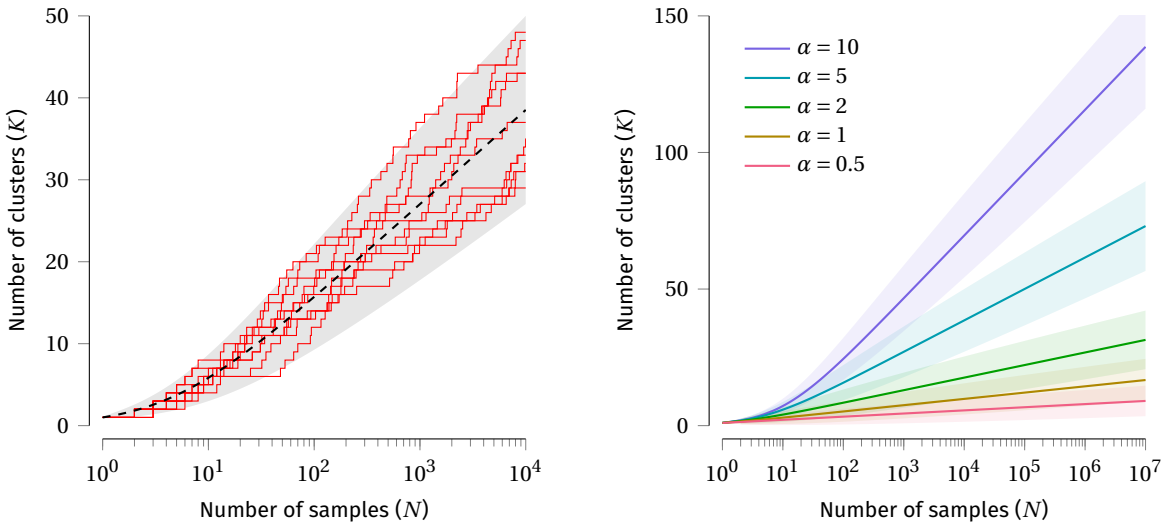
$$\theta_k^* \sim H, \qquad k = 1, \ldots, \infty, \tag{2.8a}$$

$$z_n \,|\, \mathbf{Z}_{1:n-1} \sim \text{CRP}(\alpha, \mathbf{Z}_{1:n-1}), \qquad n = 1, \ldots, N, \tag{2.8b}$$

$$\theta_n = \theta_{z_n}^*, \qquad n = 1, \ldots, N. \tag{2.8c}$$

With the above construction, $G \sim \text{DP}(\alpha, H)$ and $\theta_n \overset{iid}{\sim} G$.

The expected number of clusters generated by a CRP depends on the probability of each sample



a Traces of 10 realisations of a CRP with $\alpha = 5$

b Expected number of clusters for varying α

Figure 2.3 Behaviour of the number of DP clusters as the number of samples increases. The shaded areas represent a range of two standard deviations above and below the expected number of 'tables' in a Chinese restaurant process (CRP) with concentration α.

being assigned to a new cluster:

$$\mathbb{E}[K \mid N] = \sum_{n=1}^{N} \frac{\alpha}{\alpha + i - 1} = \alpha(\psi(\alpha + N) - \psi(\alpha)) \quad \in \quad \mathcal{O}(\alpha \log N), \tag{2.9}$$

where $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \log \Gamma(x)$ is the digamma function (Teh, 2010). The asymptotic behaviour of the number of clusters is displayed in Fig. 2.3b. In medical imaging applications such as the ones analysed in Chapter 4, N is often the number of voxels, and is typically in the 10^5–10^7 range, wherein the choice of α clearly plays a significant role in the expected complexity of the models.

Stick-breaking construction There is yet another interpretation of DPs that provides further insight into its properties: the stick-breaking process (SBP), based on a different analogy. Imagine we have a unit-length stick, and we break off a piece of length β_1. We then break the remainder of the stick at a fraction β_2 of its current length, $1 - \beta_1$, then at β_3 of $(1 - \beta_2)(1 - \beta_1)$, and so on. This process is repeated to infinity, and π_k denotes the length of the k-th stick piece (Fig. 2.4).

If the proportions β_k are drawn from a Beta$(1, \alpha)$ distribution, with the same concentration parameter α as discussed so far, the SBP can be summarised as

$$\beta_k \sim \mathrm{Beta}(1, \alpha), \qquad \pi_k = \beta_k \prod_{l=1}^{k-1} (1 - \beta_l), \qquad k = 1, \ldots, \infty, \tag{2.10}$$

and the resulting distribution over $\boldsymbol{\pi}$ is often denoted $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$, after Griffiths–Engen–McCloskey

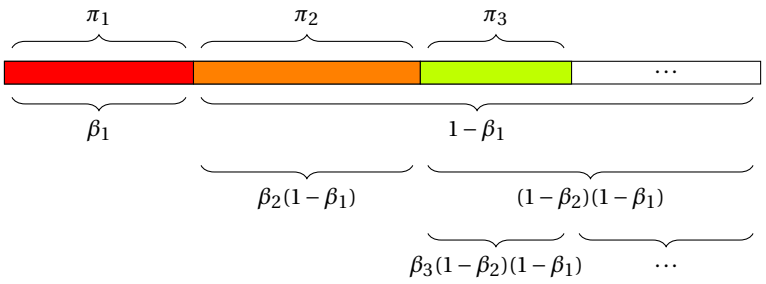
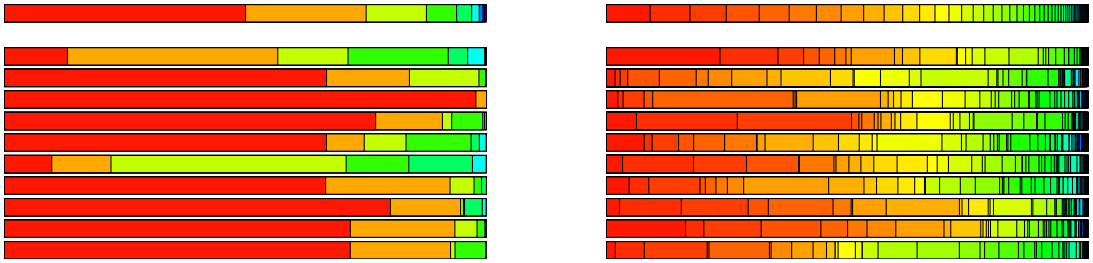


Figure 2.4 Illustration of the stick-breaking process. A stick of length 1 is successively broken into parts of length $\pi_k = \beta_k \times$ (remaining length), with $k = 1, \ldots, \infty$. In a Dirichlet process with concentration α, the break proportions are distributed as $\beta_k \sim \text{Beta}(1, \alpha)$, and we denote $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$.



a Concentration $\alpha = 1$ **b** Concentration $\alpha = 10$

Figure 2.5 Behaviour of stick-breaking processes with different concentrations. The topmost row shows the expected pattern, $\mathbb{E}[\boldsymbol{\pi}]$, and the remaining rows display different SBP realisations, $\boldsymbol{\pi} \sim \text{GEM}(\alpha)$.

(Pitman, 2006).[9] Additionally, it follows that $\sum_k \pi_k \overset{\text{a.s.}}{=} 1$, making $\boldsymbol{\pi}$ a probability mass function. Sethuraman (1994) showed that, when

$$\boldsymbol{\pi} \sim \text{GEM}(\alpha), \qquad \theta_k^* \sim H, \quad k = 1, \ldots, \infty, \qquad G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}, \tag{2.11}$$

the resulting distribution G follows a $\text{DP}(\alpha, H)$.

As can be seen in Fig. 2.5, a low value of the concentration parameter will tend to produce large and uneven pieces, whereas a higher value will generate smaller and more balanced ones. This effect is key when specifying the prior for the mixture model, because it corresponds to the level of fragmentation we expect to find in the dataset.

It bears pointing out that $\boldsymbol{\pi}$ has an infinite number of components and requires sampling an infinite number of β_k variables, which makes this representation computationally impractical. One would typically define a truncation level K and set $\beta_K = 1$ (i.e. stop breaking the stick at K), which makes $\pi_k = 0$ for $k > K$. Although this turns it into a finite model, the truncation can be a good enough approximation for large K (Ishwaran and Zarepour, 2000), because the π_k's decay exponentially on average.[10,11]

[9] This is a special case of the generalised Dirichlet distribution, where each proportion β_k follows $\text{Beta}(a_k, b_k)$. The original definition and many general properties are presented by Connor and Mosimann (1969).

[10] The expected size of each piece is $\mathbb{E}[\pi_k] = \frac{1}{1+\alpha}\left(\frac{\alpha}{1+\alpha}\right)^{k-1}$.

[11] There exists a generalisation of the stick-breaking process, called Poisson–Dirichlet or Pitman–Yor process (PYP; Pitman

Finally, the posterior distribution of $\boldsymbol{\beta}$ given a set of categorical samples $\mathbf{Z}$, with $z_n \sim \text{Cat}(\boldsymbol{\pi})$, is given by

$$\beta_k \,|\, \mathbf{Z} \sim \text{Beta}(1 + N_k, \alpha + N_{>k}), \qquad k = 1, \ldots, \infty, \tag{2.12}$$

where $N_{>k} = \sum_{l>k} N_l$. This is simply the posterior for a Beta distribution, counting the samples that fell on each side of the k-th break point.

2.2.1.2 Dirichlet process mixtures

The stick-breaking construction makes it evident that DPs, despite their flexibility, generate only discrete distributions (almost surely). Most real-world data, however, can be presumed to have continuous densities. In order to construct continuous nonparametric models based on the DP, we can use it as a prior on mixture models (Antoniak, 1974; Ferguson, 1983). As a concrete example, recall from the opening of Section 2.2.1 that we may wish to apply these models to image pixel intensities (as in Chapter 4).

In classic mixture models, we assume our observations $\mathbf{x}$ are generated by a finite mixture of parametric distributions F indexed by θ_k, with mixing proportions π_k: $p(\mathbf{x}_n \,|\, \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k F(\mathbf{x}_n \,|\, \theta_k)$. As a basic example, consider a simple Gaussian mixture model, where $\theta_k = \{\mu_k\}$ and $F(\theta_k) = \mathcal{N}(\mu_k, \sigma^2)$.

In contrast, in a Dirichlet process mixture model (DPMM), each observation is associated with the parameters of its own distribution, which are drawn i.i.d. from a DP:

$$G \sim \text{DP}(\alpha, H)\,, \tag{2.13a}$$

$$\theta_n \,|\, G \sim G, \qquad n = 1, \ldots, N\,, \tag{2.13b}$$

$$\mathbf{x}_n \,|\, \theta_n \sim F(\theta_n), \qquad n = 1, \ldots, N\,. \tag{2.13c}$$

Taking into account the clustering property of DPs, some θ_n's will have repeated values, so it is clear that the corresponding clusters of observations will have been generated by the same distributions. Rewriting Eq. (2.13) in the stick-breaking representation (Eq. (2.11)), we have

$$p(\mathbf{x}_n \,|\, G) = \sum_{k=1}^{\infty} \pi_k F(\mathbf{x}_n \,|\, \theta_k^*)\,. \tag{2.14}$$

We see that the marginal likelihood of $\mathbf{x}_n$ is an *infinite* mixture of distributions. In fact, DPMMs arise naturally as the limit in finite mixture models (Rasmussen, 2000). Just as a DP defines a prior on discrete distributions over Θ, a DPMM defines a prior on all F-mixtures, with any number of

and Yor, 1997), which is able to model power-law distributions and subsumes the DP as a special case.

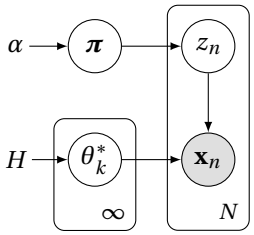


Figure 2.6 Plate diagram for a Dirichlet process mixture model. Circled nodes are random variables; uncircled nodes are fixed hyperparameters; shaded and unshaded nodes are resp. observed and unobserved variables; the rectangles ('plates') indicate conditionally independent replications of the contained nodes. Here, H is a prescribed prior distribution over component parameters $\{\theta^*_k\}_k$; α is the concentration of the GEM distribution over the mixing proportions $\boldsymbol{\pi}$; z_n is the cluster indicator; $\mathbf{x}_n$ is an observed datum, distributed as $F(\theta^*_{z_n})$.

components. With this formulation, H defines the prior on component parameters, $\{\theta^*_k\}_k$, and GEM(α) is a prior on the infinite mixing proportions, $\{\pi_k\}_k$—whereas the mixing weights are typically given a Dirichlet prior in a finite mixture.

Finally, the full generative model with indicator variables can be expressed in terms of the stick-breaking construction (Fig. 2.6):

$$\theta^*_k \sim H, \qquad k = 1, \ldots, \infty, \tag{2.15a}$$

$$\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha), \tag{2.15b}$$

$$z_n \,|\, \boldsymbol{\pi} \sim \mathrm{Cat}(\boldsymbol{\pi}), \qquad n = 1, \ldots, N, \tag{2.15c}$$

$$\mathbf{x}_n \,|\, z_n, \boldsymbol{\theta} \sim F(\theta^*_{z_n}), \qquad n = 1, \ldots, N. \tag{2.15d}$$

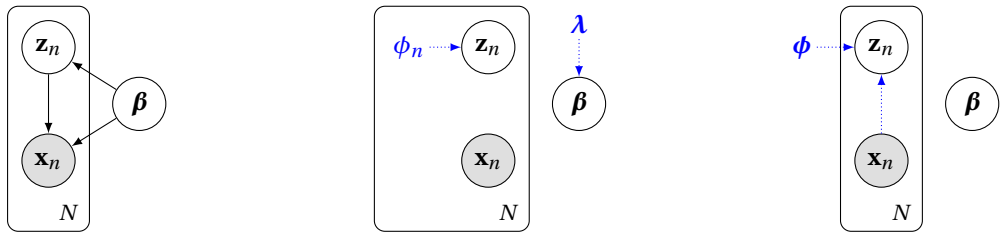
The distribution of $\mathbf{x}_n \,|\, \boldsymbol{\pi}, \boldsymbol{\theta}$ provides a density over the whole observation space, $\mathcal{X}$. If we consider additionally the indicator variables z_n (Section 2.2.1.1), the DPMM describes a full generative model that can be used for clustering the data. It constitutes a nonparametric alternative to well-established clustering methods such as k-means and finite Gaussian mixture models.

2.2.2 Variational inference

In many applications, the data-generating process can be represented by a general class of generative models consisting of observations along with local and global hidden variables (M. D. Hoffman et al., 2013). The observations, $\mathbf{X} = (\mathbf{x}_n)_{n=1}^N$, are assumed to be governed by corresponding local latent variables, $\mathbf{Z} = (\mathbf{z}_n)_{n=1}^N$, and the global latent variable $\boldsymbol{\beta}$ (possibly a collection) determines the dependence structure between observations (e.g. clustering, temporal, spatial). This is illustrated as a plate diagram in Fig. 2.7a, which reflects the factorisation of their joint probability distribution:

$$p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}) = p(\boldsymbol{\beta}) \prod_{n=1}^{N} \left[p(\mathbf{x}_n \,|\, \mathbf{z}_n, \boldsymbol{\beta})\, p(\mathbf{z}_n \,|\, \boldsymbol{\beta}) \right]. \tag{2.16}$$

As an example, a mixture model like the one described in Section 2.2.1.2 and used in Chapter 4 fits into this template: $\boldsymbol{\beta}$ corresponds to the mixing proportions $\boldsymbol{\pi}$ and component parameters $(\theta^*_k)_{k=1}^K$ (e.g. means and covariance matrices, for Gaussian mixtures), while z_n is the discrete cluster membership



a Generative model **b** Mean-field posterior **c** Amortised posterior

Figure 2.7 Plate diagrams for a general class of probabilistic models suited to variational inference. The set of global latent variables is denoted by $\boldsymbol{\beta}$, and there are local latent variables $\mathbf{z}_n$ for each observation $\mathbf{x}_n$. The assumed forward generative model is shown in **a**, while the two approaches to approximate $p(\mathbf{Z}, \boldsymbol{\beta} \mid \mathbf{X})$ described in Sections 2.2.2.1 and 2.2.2.2 are illustrated in **b** and **c**, respectively.

indicator for each observation, $\mathbf{x}_n$ (cf. Eq. (2.15) and Fig. 2.6).

Given a dataset of observations, $\mathbf{X}$, the goal of the analysis is to infer the posterior distribution over the latent variables:

$$p(\mathbf{Z}, \boldsymbol{\beta} \mid \mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta})}{\iint p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta}) \, \mathrm{d}\mathbf{Z} \, \mathrm{d}\boldsymbol{\beta}} . \tag{2.17}$$

Unfortunately, except for the very special class of conjugate models, this posterior distribution cannot be characterised analytically due to the intractable denominator. In many cases, it is infeasible even to generate samples from it.

As a way to circumvent this difficulty, variational inference proposes to approximate the posterior by a simpler distribution $q(\mathbf{Z}, \boldsymbol{\beta} \mid \mathbf{X})$ whose density we can evaluate and from which are able to sample. This is achieved by first defining a lower bound on the marginal log-likelihood of the data:

$$\begin{aligned}
\log p(\mathbf{X}) &= \log \mathbb{E}_{p(\mathbf{Z}, \boldsymbol{\beta})} [p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\beta})] \\
&= \log \mathbb{E}_{q(\mathbf{Z}, \boldsymbol{\beta} \mid \mathbf{X})} \left[p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\beta}) \frac{p(\mathbf{Z}, \boldsymbol{\beta})}{q(\mathbf{Z}, \boldsymbol{\beta} \mid \mathbf{X})} \right] \\
&\geq \mathbb{E}_{q(\mathbf{Z}, \boldsymbol{\beta} \mid \mathbf{X})} \left[\log \left(p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\beta}) \frac{p(\mathbf{Z}, \boldsymbol{\beta})}{q(\mathbf{Z}, \boldsymbol{\beta} \mid \mathbf{X})} \right) \right] \\
&= \mathbb{E}_{q(\mathbf{Z}, \boldsymbol{\beta} \mid \mathbf{X})} [\log p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\beta})] - D_{\mathrm{KL}} [q(\mathbf{Z}, \boldsymbol{\beta} \mid \mathbf{X}) \| \, p(\mathbf{Z}, \boldsymbol{\beta})] \\
&= \mathbb{E}_{q(\mathbf{Z}, \boldsymbol{\beta} \mid \mathbf{X})} \left[\sum_{n=1}^{N} \log p(\mathbf{x}_n \mid \mathbf{z}_n, \boldsymbol{\beta}) \right] - D_{\mathrm{KL}} [q(\mathbf{Z}, \boldsymbol{\beta} \mid \mathbf{X}) \| \, p(\mathbf{Z}, \boldsymbol{\beta})] =: \mathcal{L}_{\mathrm{ELBO}}(\mathbf{X}, p, q) .
\end{aligned} \tag{2.18}$$

The quantity in Eq. (2.18) is known as the *evidence lower bound* (ELBO). Variational inference proceeds by searching in some space $\mathcal{Q}$ of distributions over the latent variables, and attempts to find the member $q^* \in \mathcal{Q}$ that maximises the ELBO. Because optimisation in a free-form functional space would generally be intractable, one normally constrains $\mathcal{Q}$ to be a parametric family of distributions, i.e. $\mathcal{Q} = \{q_\phi : \phi \in \Phi\}$. This turns inference into a conventional optimisation problem over a now finite-dimensional parameter space, Φ (Bishop, 2006, Ch. 10).

2.2.2.1 Mean-field inference

The simplest class of variational distributions over the local and global latent variables is one that factorises completely and is not conditioned on the observations:

$$q_{\mathrm{MF}}(\mathbf{Z}, \boldsymbol{\beta} \,|\, \mathbf{X}) = q(\boldsymbol{\beta}) \prod_{n=1}^{N} q(\mathbf{z}_n)\,. \tag{2.19}$$

The distributions $q(\boldsymbol{\beta})$ and each of $\{q(\mathbf{z}_n)\}_{n=1}^N$ are independent and have their own parameters, resp. $\boldsymbol{\lambda}$ and $\{\phi_n\}_{n=1}^N$, that we aim to optimise (see Fig. 2.7b). This is among the most common approaches adopted for variational inference in conjugate graphical models, and is what we employ for fitting the parameters of DP Gaussian mixtures to image histograms in Chapter 4. Note that this can be a very restrictive approximation to the true posterior distribution in some cases, as it completely neglects interactions between the latent variables.

Under this factorisation, the ELBO in Eq. (2.18) becomes

$$\mathcal{L}_{\mathrm{ELBO}}(\mathbf{X}, p, q_{\mathrm{MF}}) = \mathbb{E}_{q(\boldsymbol{\beta})}\left[\sum_{n=1}^{N}\Big(\mathbb{E}_{q(\mathbf{z}_n)}[\log p(\mathbf{x}_n \,|\, \mathbf{z}_n, \boldsymbol{\beta})] - D_{\mathrm{KL}}[q(\mathbf{z}_n) \,\|\, p(\mathbf{z}_n \,|\, \boldsymbol{\beta})]\Big)\right] - D_{\mathrm{KL}}[q(\boldsymbol{\beta}) \,\|\, p(\boldsymbol{\beta})]\,. \tag{2.20}$$

This form of the optimisation objective lends itself to a coordinate ascent algorithm, whereby we iteratively maximise Eq. (2.20) with respect to each of $\{q(\mathbf{z}_n)\}_{n=1}^N$ and $q(\boldsymbol{\beta})$ in turn. Using variational calculus, it can be shown that the optimal update rule for the distribution of each latent variable is given by a partial expectation over all the other factors, which remain fixed at each step (Bishop, 2006, Sec. 10.1.1). Specifically, for the class of models considered here, we obtain

$$\log q^*(\mathbf{z}_n) = \mathbb{E}_{q_{\mathrm{MF}}(\mathbf{Z}_{-n}, \boldsymbol{\beta})}[\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta})] + \text{const.} = \mathbb{E}_{q(\boldsymbol{\beta})}[\log p(\mathbf{x}_n, \mathbf{z}_n \,|\, \boldsymbol{\beta})] + \text{const.} \tag{2.21a}$$

$$\log q^*(\boldsymbol{\beta}) = \mathbb{E}_{q_{\mathrm{MF}}(\mathbf{Z})}[\log p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\beta})] + \text{const.} = \sum_{n=1}^{N} \mathbb{E}_{q(\mathbf{z}_n)}[\log p(\mathbf{x}_n, \mathbf{z}_n \,|\, \boldsymbol{\beta})] + \log p(\boldsymbol{\beta}) + \text{const.}\,, \tag{2.21b}$$

where 'const.' on each line refers to terms that do not depend on $\mathbf{z}_n$ or $\boldsymbol{\beta}$, respectively.

When the model employs conjugate distributions and the mean-field variational factors are in the same exponential family as the model, the updates above can be computed in closed form (M. D. Hoffman et al., 2013). The resulting expressions typically correspond to the log-densities of known exponential-family distributions, and the new values of the parameters can be determined by inspection. This is the case for the approach adopted in Chapter 4, where we derive explicit update rules for the parameters of the variational distributions for each global and local latent variable.

2.2.2.2 Amortised inference

The mean-field approach fits an unconditional distribution to each local latent variable, $q(\mathbf{z}_n)$, which not only may be computationally impractical if N is large but also does not facilitate inference of $\mathbf{z}_{N+1}$ for a new observation $\mathbf{x}_{N+1}$. Moreover, the learning procedure described in the previous section relies on the use of conjugate exponential-family distributions. This modelling constraint may be too restrictive for applications involving structured high-dimensional data such as images. As we discuss in Chapter 7, these cases generally call for flexible, highly non-linear models, which in turn require more powerful inference methods.

Alternatively, we may posit a variational distribution that is conditional on observations: $q(\mathbf{z}_n \mid \mathbf{x}_n)$, called a *recognition model*. Because this can be formulated as a single function mapping from observations to distributions over the local latent variables—therefore sharing statistical strength—this formulation is called *amortised* variational inference, and can then also be applied to unseen observations in a single pass (Rezende et al., 2014).

For simplicity, let us consider the case with no global latent variables $\boldsymbol{\beta}$, or where they can be considered deterministic parameters.[12] Here, the variational posterior decomposes as

$$q(\mathbf{Z} \mid \mathbf{X}) = \prod_{n=1}^{N} q_{\boldsymbol{\phi}}(\mathbf{z}_n \mid \mathbf{x}_n)\,, \tag{2.22}$$

where now all $\{q_{\boldsymbol{\phi}}(\mathbf{z}_n \mid \mathbf{x}_n)\}_{n=1}^{N}$ share the same set of parameters, $\boldsymbol{\phi}$ (see Fig. 2.7c).

If we choose to represent $q_{\boldsymbol{\phi}}(\mathbf{z}_n \mid \mathbf{x}_n)$ as a complex non-linear function such as a neural network, its parameters can no longer be fitted by neat closed-form updates as in Section 2.2.2.1. Instead, an effective training strategy relies on (stochastic) gradient-based optimisation, the workhorse of modern deep learning. Specifically, for this we must compute the gradients of the ELBO,

$$\begin{aligned}\mathcal{L}_{\text{ELBO}}(\mathbf{X}, p, q) &= \sum_{n=1}^{N} \Big(\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_n \mid \mathbf{x}_n)} [\log p_{\boldsymbol{\theta}}(\mathbf{x}_n \mid \mathbf{z}_n)] - D_{\text{KL}}[q_{\boldsymbol{\phi}}(\mathbf{z}_n \mid \mathbf{x}_n) \,\|\, p(\mathbf{z}_n)] \Big) \\ &= \sum_{n=1}^{N} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_n \mid \mathbf{x}_n)} \left[\log p_{\boldsymbol{\theta}}(\mathbf{x}_n \mid \mathbf{z}_n) - \log \frac{q_{\boldsymbol{\phi}}(\mathbf{z}_n \mid \mathbf{x}_n)}{p(\mathbf{z}_n)} \right],\end{aligned} \tag{2.23}$$

with respect to the parameters of the recognition model, $\boldsymbol{\phi}$, but also of the generative model, $\boldsymbol{\theta}$. Although estimating derivatives of expectations is a highly active area of research (Mohamed et al., 2020), here we focus on estimators based on the *reparametrisation trick*, described below.

Proposition 2.2 (Reparametrisation trick). *Let $q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x})$ be a conditional distribution with a differenti-*

[12] There also exist more sophisticated approaches that combine non-conjugate generative and recognition models (e.g. $p_{\boldsymbol{\theta}}(\mathbf{x}_n \mid \mathbf{z}_n)$ and $q_{\boldsymbol{\phi}}(\mathbf{z}_n \mid \mathbf{x}_n)$ parametrised by deep neural networks) with conventional conjugate priors on the global variables $\boldsymbol{\beta}$ (e.g. Johnson et al., 2016; Lin et al., 2018). Complementarily, the sub-field of Bayesian deep learning investigates tractable approaches to handling neural networks whose weights are themselves random variables.

able non-centred parametrisation (Kingma and Welling, 2014b), i.e. there exist a differentiable function $\mathbf{g}_{\boldsymbol{\phi}}$ *and a base distribution* $p(\boldsymbol{\epsilon})$ *independent of* $\boldsymbol{\phi}$ *such that* $\mathbf{z} = \mathbf{g}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{x}) \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$*. The expectation of a differentiable function* f *can then be approximated via ancestral sampling and Monte Carlo:*

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})] = \mathbb{E}_{p(\boldsymbol{\epsilon})}[f(\mathbf{g}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{x}))] \approx \frac{1}{S}\sum_{s=1}^{S} f(\mathbf{g}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(s)}, \mathbf{x})), \tag{2.24}$$

where $\boldsymbol{\epsilon}^{(s)} \sim p(\boldsymbol{\epsilon})$ *for* $s = 1, \ldots, S$*. Similarly, one can estimate the gradient of this expectation w.r.t. the parameters* $\boldsymbol{\phi}$ *as (Kingma and Welling, 2014a; Rezende et al., 2014)*

$$\begin{aligned}\nabla_{\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[f(\mathbf{z})] &= \nabla_{\boldsymbol{\phi}} \mathbb{E}_{p(\boldsymbol{\epsilon})}[f(\mathbf{g}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{x}))] = \mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla_{\boldsymbol{\phi}} f(\mathbf{g}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{x}))] \\ &= \mathbb{E}_{p(\boldsymbol{\epsilon})}[\nabla_{\mathbf{z}} f(\mathbf{g}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{x})) \cdot \nabla_{\boldsymbol{\phi}} \mathbf{g}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{x})] \\ &\approx \frac{1}{S}\sum_{s=1}^{S} \nabla_{\mathbf{z}} f(\mathbf{g}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(s)}, \mathbf{x})) \cdot \nabla_{\boldsymbol{\phi}} \mathbf{g}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}^{(s)}, \mathbf{x}).\end{aligned} \tag{2.25}$$

In particular, the most widespread and successful implementation of this framework is the *variational auto-encoder* (VAE; Kingma and Welling, 2014a). This algorithm models the generative and recognition distributions, $p_{\boldsymbol{\theta}}(\mathbf{x}_n|\mathbf{z}_n)$ and $q_{\boldsymbol{\phi}}(\mathbf{z}_n|\mathbf{x}_n)$, by means of neural networks that output the parameters of simple distributions over $\mathbf{x}$ and $\mathbf{z}$, respectively—which can thus be seen as a pair of probabilistic decoder and encoder.

The archetypal encoder predicts the mean and dimension-wise variances of an axis-aligned multivariate normal distribution over $\mathbf{z}$, i.e. $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})))$ (Kingma and Welling, 2014a). The corresponding reparametrisation is given by

$$\mathbf{z} = \mathbf{g}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{x}) = \boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}) + \boldsymbol{\sigma}_{\boldsymbol{\phi}}(\mathbf{x}) \odot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2.26}$$

Symmetrically, the decoder typically predicts the parameters of simple distributions over $\mathbf{x}$, e.g. pixel means and variances for images. This is the formulation adopted in our experiments in Chapter 7.

2.2.3 Flow-based models

A powerful alternative approach to probabilistic modelling is based on invertible transformations between random variables. Its key advantage is that, when the source variable's density is known, the likelihood of the transformed variable can be computed exactly.

Proposition 2.3 (Change of variables). *Assume each observation* $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^D$ *has an associated local latent variable* $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^D$ *of same dimensionality, and that they are related via a diffeomorphic trans-*

formation[13] $\mathbf{g}: \mathcal{Z} \to \mathcal{X}$. *If* $\mathbf{x} = \mathbf{g}(\mathbf{z})$, *with* $\mathbf{z} \sim p(\mathbf{z})$, *then the probability density function of* $\mathbf{x}$ *can be evaluated via the* change of variables formula *(also known as the Jacobian equation):*

$$p(\mathbf{x}) = p(\mathbf{z})\big|_{\mathbf{z}=\mathbf{g}^{-1}(\mathbf{x})} \cdot \left|\det \nabla \mathbf{g}^{-1}(\mathbf{x})\right| = \left.\frac{p(\mathbf{z})}{\left|\det \nabla \mathbf{g}(\mathbf{z})\right|}\right|_{\mathbf{z}=\mathbf{g}^{-1}(\mathbf{x})}, \tag{2.27}$$

where the second equality follows from the inverse function theorem.

The term $\nabla \mathbf{g}(\mathbf{z}) \in \mathbb{R}^{D \times D}$, known as the *Jacobian* of $\mathbf{g}$, is the matrix of its partial derivatives, such that $[\nabla \mathbf{g}]_{ij} = \partial g_i / \partial z_j$. The absolute Jacobian determinant can be seen as a factor that 'corrects' for distortions in volume introduced by the transformation $\mathbf{g}$.[14] A value of $|\det \nabla \mathbf{g}| > 1$ indicates an expansion, such that the density $p(\mathbf{x})$ must be adjusted downward; conversely, there is a compression when $|\det \nabla \mathbf{g}| < 1$, and the density increases.

When $p(\mathbf{z})$ is a known base density that can be easily evaluated and $\mathbf{g}$ is a flexible nonlinear function, this relation can be used to build expressive probabilistic models of observations $\mathbf{x}$. In the most typical usage in machine-learning literature, $p(\mathbf{z})$ is taken to be a simple, fixed distribution such as the standard (multivariate) Gaussian, while $\mathbf{g}$ is a function whose parameters are fitted to the data via maximum-likelihood (Tabak and Vanden-Eijnden, 2010). Such models are called *normalising flows*, as the inverse transformation $\mathbf{g}^{-1}$ maps a complex, unknown distribution, $p(\mathbf{x})$, to the prescribed one, $p(\mathbf{z})$ (incidentally, often a normal distribution). More details and comprehensive literature reviews of recent flow-based models can be found in Kobyzev et al. (2020) and Papamakarios et al. (2019).

In the following sections we present an overview of two approaches for leveraging flows in probabilistic models: as a stack of simple computational layers, which we later employ for modelling distributions of non-imaging covariates in Chapter 7; or as smooth deformations, a concept we exploit for image intensity normalisation in Chapter 4.

2.2.3.1 Compositional flows

While normalising flows are conceptually simple, they come with notable computational constraints. In particular, the cost of calculating the Jacobian determinant $\det \nabla \mathbf{g}$ generally scales cubically with the dimensionality D,[15] making it impractical for high-dimensional applications. A judicious choice of the functional form of $\mathbf{g}$ may however circumvent this limitation, by ensuring its Jacobian $\nabla \mathbf{g}$ is more computationally convenient (e.g. diagonal, triangular, block structure).

On the other hand, such a choice may severely restrict the expressiveness of the transformation.

[13] Recall that a diffeomorphism is a differentiable and invertible function whose inverse is also differentiable.
[14] Because $\mathbf{g}$ is diffeomorphic, its Jacobian determinant $\det \nabla \mathbf{g}$ is non-zero everywhere and has constant sign.
[15] The cubic cost comes from first applying an LU decomposition to the Jacobian matrix; the naïve computation via recursive cofactor expansion scales instead as $\mathcal{O}(D!)$.

Fortunately, a composition of diffeomorphisms is also diffeomorphic. In analogy with conventional deep neural networks, this means we can construct a more elaborate normalising flow by stacking L simpler transformations $(\mathbf{g}_l)_{l=1}^L$ with tractable Jacobian determinants (Tabak and C. V. Turner, 2013):

$$\mathbf{g} = \mathbf{g}_L \circ \cdots \circ \mathbf{g}_2 \circ \mathbf{g}_1 \,. \tag{2.28}$$

Now, invoking the chain rule of derivatives and the multiplication property of determinants, we can show that the Jacobian determinant of $\mathbf{g}$ is the product of the Jacobian determinants of each elementary layer:

$$\det \nabla \mathbf{g}(\mathbf{z}_0) = \det\big[\nabla \mathbf{g}_L(\mathbf{z}_{L-1}) \cdots \nabla \mathbf{g}_2(\mathbf{z}_1) \cdot \nabla \mathbf{g}_1(\mathbf{z}_0)\big] = \prod_{l=1}^{L} \det \nabla \mathbf{g}_l(\mathbf{z}_{l-1}) \,, \tag{2.29}$$

where we denote the intermediate outputs $\mathbf{z}_l = \mathbf{g}_l(\mathbf{z}_{l-1})$, with the initial value sampled from the base distribution, $\mathbf{z}_0 \sim p(\mathbf{z})$.

Finally, combining Proposition 2.3 with Eq. (2.29), the marginal log-likelihood $\log p(\mathbf{x})$ of a datum $\mathbf{x}$ can be computed as

$$\log p(\mathbf{x}) = \log p(\mathbf{z}) - \sum_{l=1}^{L} \log\big|\det \nabla \mathbf{g}_l(\mathbf{z}_{l-1})\big| \tag{2.30a}$$

$$= \log p(\mathbf{z}) + \sum_{l=1}^{L} \log\big|\det \nabla \mathbf{g}_l^{-1}(\mathbf{z}_l)\big| \,, \tag{2.30b}$$

where $\mathbf{z}_{l-1} = \mathbf{g}_l^{-1}(\mathbf{z}_l)$ and we define $\mathbf{z}_L := \mathbf{x}$ and $\mathbf{z}_0 := \mathbf{z}$ for notational convenience. Either form (2.30a) or (2.30b) may be used, depending on whether the forward or backward Jacobian determinants—$(\det \nabla \mathbf{g}_l)_{l=1}^L$ or $(\det \nabla \mathbf{g}_l^{-1})_{l=1}^L$, respectively—are more convenient to compute.

The growing literature on normalising flows presents a wide variety of elementary transformations $\mathbf{g}_l$ suitable for building powerful flow-based models. Specifically, in Chapter 7 we employ first-order spline flows (Durkan et al., 2019) to model unconditional distributions over scalar covariates and represent conditional distributions using affine layers (Lu and B. Huang, 2020). More details will be presented in context.

2.2.3.2 Differential flows

An alternative perspective on the change of variables considers mappings defined as a continuous composition of infinitesimal transformations, as defined by the following ordinary differential equation (ODE):

$$\frac{\partial \mathbf{g}_t}{\partial t} = \mathbf{f}_t \circ \mathbf{g}_t, \qquad \mathbf{g}_0(\mathbf{z}) = \mathbf{z}. \tag{2.31}$$

Here, t is a continuous 'time' variable that indexes the transformation from 0 up to some end time T (which can be taken as 1, without loss of generality). Borrowing from the Lagrangian perspective in fluid mechanics, the function $\mathbf{f}_t$ can be understood as a 'velocity field' that transports 'particles' $\mathbf{z}$ over time, and $(\mathbf{g}_t)_{t\in[0,T]}$ traces the corresponding trajectories. Under some regularity conditions on $\mathbf{f}_t(\mathbf{z})$, there is a unique solution $(\mathbf{g}_t(\mathbf{z}))_{t\in[0,T]}$ for a given initial condition $\mathbf{z}$.[16] When the velocity field $\mathbf{f}_t$ is represented by a neural network, the model is known as a neural ODE (R. T. Q. Chen et al., 2018b).

The result of the transformation, defined as the final position of a particle, can be determined by integrating the ODE (2.31):

$$\mathbf{x} = \mathbf{g}_T(\mathbf{z}) = \mathbf{z} + \int_0^T \mathbf{f}_t(\mathbf{g}_t(\mathbf{z}))\,\mathrm{d}t\,. \tag{2.32}$$

Though the integral above can rarely be solved analytically, there are many numerical methods available for obtaining acceptable approximations. For example, our investigation in Chapter 4 employed the classic Runge–Kutta solver (RK4), whose accumulated error rate is quartic in the discretisation step size.

Assuming $\mathbf{g}_0(\mathbf{z}) = \mathbf{z} \sim p_0(\mathbf{z})$, the change in the density can be described by the *continuity equation*:

$$\frac{\partial p_t}{\partial t} + \nabla\cdot(p_t\,\mathbf{f}_t) = 0\,. \tag{2.33}$$

This relation defines the property of *conservation of mass*, a cornerstone of fluid dynamics. Taking a probability density for the typical mechanical density, the conservation of (probability) mass principle states that the probability of a point being in a fixed region of space changes by the net probability influx through its boundary. Alternatively, stated from the Lagrangian perspective, the probability of a moving region remains constant as its boundary is transported by a velocity field. When the densities and the velocities are sufficiently smooth, both pictures are equivalent to the differential conservation law in Eq. (2.33).

In order to evaluate the density at time t, the change of variables formula (Proposition 2.3) requires calculation of the Jacobian determinant of $\mathbf{g}_t$, the transformation from an initial time t_0 up to t:

$$\det\nabla\mathbf{g}_t(\mathbf{z}) = \det\nabla\mathbf{g}_{t_0}(\mathbf{z})\cdot\exp\int_{t_0}^{t}\operatorname{tr}[\nabla\mathbf{f}_\tau(\mathbf{g}_\tau(\mathbf{z}))]\,\mathrm{d}\tau\,. \tag{2.34}$$

The resulting log-density at time T is then given by

$$\log p_T(\mathbf{x}) = \log p_0\Bigg(\underbrace{\mathbf{x} + \int_0^T -\mathbf{f}_{T-t}(\overleftarrow{\mathbf{g}}_t(\mathbf{x}))\,\mathrm{d}t}_{\mathbf{z}=\overleftarrow{\mathbf{g}}_0(\mathbf{x})=\mathbf{g}_T^{-1}(\mathbf{x})}\Bigg) + \underbrace{\int_0^T \operatorname{tr}[-\nabla\mathbf{f}_{T-t}(\overleftarrow{\mathbf{g}}_t(\mathbf{x}))]\,\mathrm{d}t}_{\log\det\nabla\overleftarrow{\mathbf{g}}_0^{-1}(\mathbf{x})=\log\det\nabla\mathbf{g}_T^{-1}(\mathbf{x})}\,, \tag{2.35}$$

[16] The Picard–Lindelöf (or Cauchy–Lipschitz) theorem assures existence and uniqueness of the solution if $\mathbf{f}_t(\mathbf{z})$ can be assumed uniformly Lipschitz continuous in 'space' $\mathbf{z}$ and continuous in 'time' t (Coddington and Levinson, 1955, Ch. 1).

where we define $\overline{\mathbf{g}}_t$ as the solution to $\partial_t \overline{\mathbf{g}}_t = -\mathbf{f}_{T-t} \circ \overline{\mathbf{g}}_t$, with $\overline{\mathbf{g}}_0(\mathbf{x}) = \mathbf{x}$. In other words, we integrate the original ODE (2.31) backward in time, from $t = T$ to $t = 0$, starting at the location $\mathbf{x}$ whose likelihood we wish to compute.

Chapter 4 leverages these principles to match image intensity distributions, modelled as nonparametric mixture densities. There, we derive a form for the velocity field $\mathbf{f}_t$ that conserves mass for the evolving density as the mixture parameters are fitted to the target distribution.

2.3 Causality

Causal theory is one of the central pillars of the work in this thesis. Section 2.3.1 introduces Pearl's ladder of causation, a useful hierarchy that characterises data analysis tasks according to their goals and necessary assumptions. Then, Section 2.3.2 explains how to interpret causal relationships, followed by a description of causal graphical models and how they differ from conventional probabilistic models in Section 2.3.3. Lastly, Section 2.3.4 gives some advice on how to construct causal models from scratch for new application domains.

2.3.1 The causal hierarchy

We begin by building some intuition on the types of questions that can only be answered via causality and what assumptions are required to address them. Generally, queries about data can be classified as 'seeing', 'doing', or 'imagining', in what Pearl calls the 'Ladder of Causation' (Pearl and Mackenzie, 2018), otherwise referred to as the causal hierarchy (Bareinboim et al., 2020; Pearl, 2019). Bareinboim et al. (2020) recently showed mathematically that the levels are indeed distinct and their capabilities are strictly increasing, almost surely (in a measure-theoretical sense).[17]

Associational ('seeing') The bottom of the hierarchy describes reasoning about passively observed data, such that its main object of interest is the observational distribution produced by the system under study. Here, the goal is to draw inferences about marginal, joint, and conditional probability distributions over the various measurements collected in a dataset, addressing questions of the type *"What are the odds that I observe…?"*. It is at this level that most types of machine-learning algorithms operate (e.g. the ones presented in Section 2.2), building predictive or generative models to capture correlation patterns (statistical dependence) found in the data.

[17] In simple terms, each level is strictly less powerful than the next, except for virtually impossible coincidences.

Interventional ('doing') The next stage aims to predict the effects of interactions with the environment: *"What happens if I do…?"*. This kind of estimation problem is known as *causal inference*. Interventions are arguably at the heart of scientific enquiry, whereby one typically formulates experiments to ascertain the causal effects between variables. We attempt to isolate the mechanism of interest by controlling as many extraneous factors as necessary, and measure the target outcome after deliberately manipulating the putative cause. For example, a randomised controlled trial (RCT) assigns a treatment at random to each participant, independently of their profile and of whether they would naturally have been prescribed said treatment, to then verify if it has the desired effect on the outcome across the study cohort.

In many cases, it is possible to perform causal inference even without explicitly carrying out experiments, based only on the naturally occurring observational distribution. The formalism known as *do*-calculus is able not only to establish how to estimate the desired causal quantities from observational data, but also to determine when that is impossible (Pearl, 2009, Sec. 3.4). This procedure requires strictly more information than the association level of the hierarchy, in the form of structural assumptions: we must leverage domain knowledge to trace the presumed causal relationships underlying the data-generating process as a graph. It is this graph, the *causal diagram*, that enables predicting qualitatively how the generative process changes under interventions. We discuss the interpretation of causal diagrams in more detail in Section 2.3.2.

Counterfactual ('imagining') The last level deals with retrospective hypothetical scenarios. It is so called because inference at this level is about what would have happened—possibly *contrary to fact*—given alternative circumstances. In other words, we predict the results of imaginary retroactive interventions that may contradict the observed reality. Counterfactual reasoning is often involved in deriving explanations, namely attributing causes to observed outcomes, e.g.: Was it the medicine that cured her headache? Would this individual had gotten the job, were their skin a different colour? Would this patient have survived, had they received a different treatment? Note that such questions differ from interventional queries by referring to specific individual units, whereas the latter consider the aggregate causal effects at the population level (e.g. average treatment effects).

In addition to the observational distribution and to the graphical assumptions required by the interventional level, counterfactuals also need access to a full mechanistic description of the generative process. Because by definition counterfactuals can never be observed (i.e. there is no 'ground truth'),[18]

[18]This also means that counterfactuals are generally unfalsifiable, since the true generative model is unknown.



a Mediation **b** Confounding **c** Collider **d** Combined model

Figure 2.8 Simple causal diagrams illustrating fundamental multivariate relationships. In the example discussed in the text, A is sunlight exposure, B is damage to skin cells, C is development of skin cancer, D are vitamin D levels, and E represents genetic predisposition.

counterfactual predictions are possible only through a suitable functional model of reality—known as a *structural causal model*. Chapter 7 introduces a novel approach to high-dimensional counterfactual inference, by building structural causal models using deep-learning components such as normalising flows (Section 2.2.3) and variational auto-encoders (Section 2.2.2.2).

2.3.2 Causal reasoning

Causal reasoning is the process of analysing the data-generating process in terms of cause–effect relationships. To motivate our discussion of causality, it is useful to consider the following principle (Reichenbach, 1956; Pearl, 2009, Ch. 2; J. Peters et al., 2017, Sec. 1.3).

Proposition 2.4 (Common cause principle (Reichenbach, 1956)). *If variables A and B are observed to be correlated, then either one causes the other (possibly indirectly) or they share a common cause (possibly unobserved) that affects both of them.*

In other words, while 'correlation does not imply causation', as often repeated, this postulate states that any statistical dependence can fundamentally be explained by an underlying causal structure (J. Peters et al., 2017, Sec. 6.5.1)—'no correlation without causation'.

One can formalise causation as follows: a variable A is a direct cause of variable B, written $A \to B$, if forcing A to different values changes the likelihood of B, all else held constant (Pearl, 2009). In other words, B (say, an outcome) is assumed to have a mechanistic dependence on A (say, exposure) and potentially also on other factors and on independent noise (J. Peters et al., 2017). Crucially, $A \to B$ entails that the distribution of the cause, $P(A)$, does not inform or influence the conditional $P(B \mid A)$, a principle known as *independence of cause and mechanism* (Daniušis et al., 2010; J. Peters et al., 2017).[19]

Taking this a step further, the postulated causal links between multiple variables form a directed acyclic graph (DAG), called a causal diagram. Such graphs encode assumptions about direct and indir-

[19] The more general *principle of independent mechanisms* states that the conditional distribution of each variable given its direct causes does not inform or influence the remaining conditionals (J. Peters et al., 2017, Sec. 2.1). Intuitively, we imagine the system as a network of autonomous modules that are invariant to changes in each other. This also relates to the idea that exogenous factors (i.e. noise) affect each variable independently.

ect causal links and capture statistical information about variables such as conditional independences. To illustrate this more concretely, let us analyse some canonical relationships between three variables.

- If A (say, exposure to sunlight) affects a variable C (say, skin cancer) indirectly through its impact on B (damage to the skin cells), illustrated with the causal diagram in Fig. 2.8a, we say B is a mediator and A is an indirect cause of C. Here, B completely screens off the effect of A on C, meaning $A \perp\!\!\!\perp C \mid B$ (read 'A is conditionally independent of C, given B').

- Alternatively, assume A is a common cause of B and D (say, vitamin D levels), represented by the causal diagram in Fig. 2.8b. In this case, A is known as a confounder, producing an association between B and D, thus $B \not\perp\!\!\!\perp D$ (read 'B is not independent of D'). However, controlling for A makes them independent: $B \perp\!\!\!\perp D \mid A$.

- Finally, consider the case wherein B is a common effect of A and E (say, genetic predisposition), illustrated in Fig. 2.8c. Here, B is called a collider. Unlike the two situations above, this configuration implies A and E are independent *a priori*. On the other hand, conditioning on B introduces an association between A and E, as they may now 'explain away' the effect of each other on the observed outcome, B (i.e. $A \not\perp\!\!\!\perp E \mid B$) (Pearl, 2009).

For more general graph structures, such as the full example diagram in Fig. 2.8d, one should reason in terms of paths (i.e. chains of nodes connected by edges pointing in any direction), as they are the conduits for correlations propagated across the graph. Any path that does not contain a collider is said to be unblocked or open, and implies a potential statistical association between its endpoints.[20] Conversely, a path containing a collider is said to be blocked or closed, and does not carry any indirect causal influence between its endpoints *a priori* (Greenland and Pearl, 2011). If there are no unblocked paths between two variables, we conclude they are independent. As mentioned above, however, conditioning on a collider (or on a descendant of one) may unblock previously blocked paths.

In causality, what is usually referred to as 'bias' is any spurious correlation between two variables, contributed by unblocked paths beside the relationship of interest. One may then include certain variables as additional predictors in a model to block such undesired open paths. This is what is meant by 'adjusting/controlling for' or 'conditioning on' a variable in the contexts of predictive modelling, statistics, and causal inference.

A purely statistical perspective would be unable to fully distinguish all three configurations in Figs. 2.8a to 2.8c, making it difficult to decide what to control for. The causal perspective, on the other hand, requires us to be clear about our assumptions and immediately reveals possible confounding.

[20] Note that a path without a collider must either be directed, indicating (indirect) causation between the endpoints, or contain a confounder with directed subpaths on each side. The ensuing potential for correlation is then a manifestation of Reichenbach's principle (Proposition 2.4).

Under this model, for example, although vitamin D levels (D) are predictive of skin cell damage (B), taking vitamin supplements would be assumed to have no effect on the sun-damaged DNA molecules. The fact that causal models allow us to enquire about the effects of interventions is what sets them apart from pure statistical models, which are limited to studying correlations. This illustrates that careful considerations may be required when making decisions about data collection, sample selection, and subsequent analysis. With the ability to formalise causal concepts in clear mathematical and probabilistic terms, causal reasoning opens the door for researchers to go beyond association by allowing them to incorporate domain expertise when answering fundamental scientific questions.

2.3.3 Causal graphical models

In a causal diagram, nodes represent variables of interest and edges between them indicate postulated direct causal influences. Such a causal graphical model embodies the causal Markov assumption (or local Markov): every node is statistically independent of its non-effects (non-descendants), given its direct causes (parents). Therefore, the joint probability distribution over all variables $(V_i)_{i=1}^M$ on the graph can be factorised as a product of independent conditional mechanisms (Greenland and Pearl, 2011; J. Peters et al., 2017):

$$P(V_1, V_2, \ldots, V_M) = \prod_{i=1}^{M} P(V_i \,|\, \mathrm{pa}(V_i)) \,. \tag{2.36}$$

where $\mathrm{pa}(V_i)$ denotes the set of parents of variable V_i, i.e. the nodes with arrows pointing towards V_i. For instance, noting the graphical structure in Fig. 2.8d, the observational distribution of the example in Section 2.3.2 factorises as $P(A,B,C,D,E) = P(A)\,P(B \,|\, A,E)\,P(C \,|\, B)\,P(D \,|\, A)\,P(E)$.

For those familiar with Bayesian networks, it appears that there is nothing new. However, Bayesian networks only encode conditional independence relationships, and they are thus not unique for a given observational distribution (Pearl, 2009). In fact, although causal arguments often guide the construction of such models, any alignment between arrows in Bayesian networks and causality is merely coincidental. In particular, causal models differ from Bayesian networks in that, beside representing a valid factorisation of the joint probability distribution, they enable reasoning about *interventions* (Pearl, 2009).

In causal graphs, the values for each node are assumed to be determined via *independent mechanisms* (cf. independence of cause and mechanism, Section 2.3.2) given their direct causes. An intervention is defined as any forced change to the mechanism generating a variable. Under the simplest and most common kind of intervention, called atomic, the variable is set to a constant, regardless of the values of its natural causes. This results in a modified graph wherein this node is disconnected from its parents, though crucially all other mechanisms are untouched. From this

modified diagram, one can derive the distribution over the remaining variables—which differs from simple conditioning because the distributions of causes of an intervened variable are unaffected. Incidentally, this is the principle behind randomised controlled trials: a treatment is assigned at random (an intervention on the 'treatment' variable), isolating its direct effect on the outcome by eliminating the influence of confounding factors (i.e. cutting the edges from common causes of treatment and outcome).

2.3.4 Building a causal diagram

The first step in constructing a causal model for a given system is to elicit the relevant variables to represent, which may be observed or not. These ought to be well defined: they should unambiguously correspond to real or postulated entities of the system, and separate variables must not have overlapping meanings (Swanson and Ibrahim, 2013). In the medical imaging context, variables normally correspond to the collected data elements, such as images, meta-information fields, labels, patient records, etc. Not all important variables need to be concrete and measurable, however. Other relevant abstract concepts can be instantiated if they help in describing complex processes: e.g. 'annotation policy', 'patient's health status', 'proprietary image post-processing pipeline'.

Secondly, the causal links between the defined variables must be determined. While each added arrow between two nodes in the graph corresponds to assuming causation, it is important to consider that the *absence* of an arrow also encodes a strong assumption. Namely, that there is no direct causal effect—any marginal association between those variables is fully explained via mediator variables or common causes (see Section 2.3.2). In addition, the granularity of 'direct effects' is only relative to the chosen level of abstraction (Greenland and Pearl, 2011). One may wish to detail the complete chain of effects between two causally linked variables, or represent them by a single arrow (e.g. $A \to B_1 \to B_2 \to C$ vs. $A \to C$).

In what is called a *selection diagram* (Pearl and Bareinboim, 2014), one also includes special indicator variables that identify the 'domain' or 'environment', e.g. training vs. testing or a certain hospital in a multi-site study. Their direct causal effects (outgoing arrows) represent the specific mechanisms through which one assumes the observed populations differ, whereas the absence of a link from a domain selector to a variable implies the latter's mechanism is invariant across environments (Pearl and Bareinboim, 2014). Domain indicators should normally be represented by root nodes in the diagram, with no incoming edges, as they embody *exogenous* changes to the data distributions. A causal diagram may additionally be augmented with selection variables, when the dataset is subject to preferential subsampling from the population (e.g. inclusion criteria for a clinical trial). The incoming arrows to such a node represent the various selection criteria (deliberate

or otherwise) that impacted the collection of the dataset of interest.

Finally, note that this construction is an iterative process. Once a full version of the diagram is written, one must verify that the assumptions implied by the graph match the domain knowledge (see notes on interpretation in Section 2.3.2), and corrections should be made as needed. Further, recall the diagram's intent as a communication tool when choosing its level of abstraction, as there is often a trade-off to be made between accuracy and clarity: the graph should be sufficiently detailed not to omit relevant variables and pathways, though no more complex than necessary (Swanson and Ibrahim, 2013).

Chapter 3

Causal reasoning for medical image analysis

This chapter is based on the following publication:

- **Daniel C. Castro**, Ian Walker and Ben Glocker (2020). 'Causality matters in medical imaging'. *Nature Communications* 11, 3673. DOI: 10.1038/s41467-020-17478-w.

Causal reasoning can shed new light on some of the major challenges in machine learning for medical imaging: scarcity of high-quality annotated data and mismatch between the development dataset and the target environment. As we will see in this chapter, a causal perspective on these issues allows decisions about data collection, annotation, pre-processing, and learning strategies to be made and scrutinised more transparently, while providing a detailed categorisation of potential biases and mitigation techniques. Along with worked clinical examples, we highlight the importance of establishing the causal relationship between images and their annotations, and offer step-by-step recommendations for future studies.

3.1 Introduction

Tremendous progress has been achieved in predictive analytics for medical imaging. With the advent of powerful machine-learning approaches such as deep learning, staggering improvements in predictive accuracy have been demonstrated for applications such as computer-aided diagnosis (Esteva et al., 2017) or assisting radiotherapy planning and monitoring of disease progression via automatic contouring of anatomical structures (Menze et al., 2015). However, two of the main obstacles for translating these successes to more applications and into wider clinical practice remain: *data scarcity*, concerning the limited availability of high-quality training data required for building predictive

models; and *data mismatch*, whereby a model trained in a lab environment may fail to generalise to real-world clinical data.

Let us exemplify with a hypothetical scenario how these obstacles may arise in practice and pose real threats to the success of research projects. Suppose a team of academic radiologists is excited about the opportunities artificial intelligence seems to offer for their discipline. In a recent study, the clinical team was able to demonstrate the effectiveness of using human interpretation of magnetic resonance imaging (MRI) for diagnosis of prostate cancer, yielding higher sensitivity and specificity than a conventional diagnostic test, as confirmed via ground-truth labels from histopathology. Motivated by these results, the team decides to approach a machine-learning (ML) research lab with the idea of developing a tool for automated, MRI-based diagnosis of prostate cancer. Because reading MRI requires advanced training and experience, they hope such a system may facilitate widespread adoption of MRI as a novel, accurate, and cost-effective tool for early diagnosis, especially in locations with lower availability of the required human expertise.

The clinicians still have access to their previous study data, and are confident this may be used for ML development. Unfortunately, the sample size is small—there are insufficient pairs of images and diagnosis labels to train a state-of-the-art deep learning image classification method. However, the clinicians have access to large amounts of (unlabelled) routine MRI scans. The ML researchers are hopeful they can additionally leverage this data in a so-called semi-supervised learning strategy. After a pilot phase of development, the team is planning to evaluate their method in a large multi-centre study.

What are the chances of success for their project, and how could a causal analysis help them to identify potential issues in advance? Regarding the limited availability of annotated data, here the team may be lucky in successfully exploiting the unlabelled data thanks to the anticausal direction between images and confirmed diagnosis labels (as we will discuss later in more detail). However, major obstacles may arise due to data mismatch between the ML development and clinical validation stage, resulting from specific inclusion criteria (selection bias), varying patient populations (e.g. changes in demographics), and prevalence of disease (e.g. due to environmental factors). Identifying these issues is important for properly designing prospective validation studies. While researchers are generally aware of the adverse effects of such differences in aspects of the data, they may be unaware that causal reasoning provides tools for laying out any underlying assumptions about the data generating process in a clear and transparent fashion, such that any issues can be more easily identified beforehand and possibly resolved by employing suitable data collection, annotation, and machine-learning strategies.

In this article we discuss how causal considerations in medical imaging can shed new light on the above challenges—illustrated with cartoon examples in Fig. 3.1—and help in finding appropriate

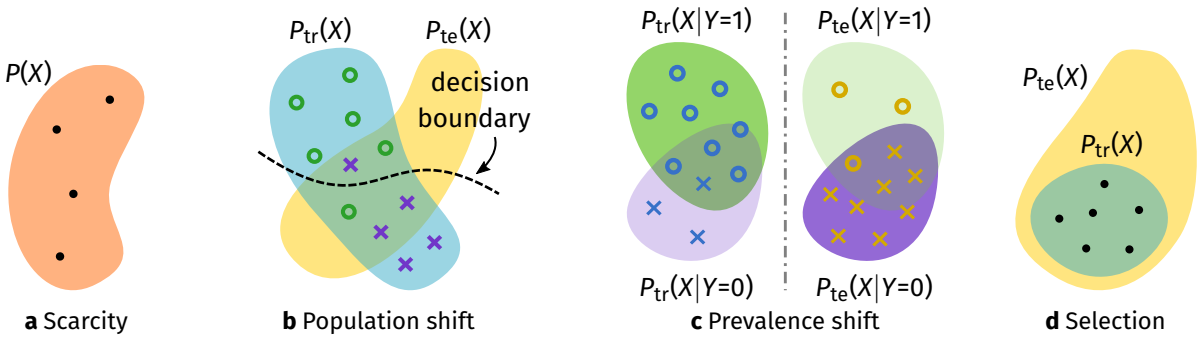


Figure 3.1 Key challenges in machine learning for medical imaging: a data scarcity and **b–d** data mismatch. X represents images and Y, annotations (e.g. diagnosis labels). P_{tr} refers to the distribution of data available for training a predictive model, and P_{te} is the test distribution, i.e. data that will be encountered once the model is deployed. Dots represent data points with any label, while circles and crosses indicate images with different labels (e.g. cases vs. controls).

solutions. In particular, we demonstrate how the causal structure of a task can have profound, and sometimes surprising, consequences on the soundness of the employed machine-learning approach and resulting analysis. We highlight that being aware of causal relationships, and related issues such as dataset shift and selection bias, allows for systematic reasoning about what strategies to prefer or to avoid. Here, the language of causal diagrams provides explicit means to specify assumptions, enabling transparent scrutiny of their plausibility and validity (Bareinboim and Pearl, 2016). It is in fact a natural way of defining the relationships between variables of interest, because it reflects the expert's knowledge of the biological and logistical processes involved in the generation and collection of data, and has been successfully applied for building models for decision-making in healthcare, for example (Cypko et al., 2017; Lucas et al., 2004). We hope our work can serve as a practical guide and inspire new directions for research in medical imaging.

3.2 Causality matters

Before diving into details of the challenges of data scarcity and data mismatch, the causal properties of the core predictive task must be analysed. In particular, one must pay close attention to the relationship between the inputs and targets of the model.

3.2.1 Predictive analytics in medical imaging

The focus of this work is on predictive modelling: given an image X, train a model to predict some given annotation Y. Specifically, we wish to estimate the conditional probability distribution $P(Y \mid X)$ by fitting a statistical model with a suitable objective function. This formulation encompasses a variety of common medical image analysis tasks, such as semantic segmentation (i.e. contouring of structures

of interest), disease classification, outcome prediction, and many more.

In this context, it is worth clarifying some terminology regarding the data that is used for development and after deployment, in order to avoid confusion of some terms that are sometimes used differently in clinical and machine-learning communities. We refer to an annotated dataset with pairs (X, Y) as the development data, which is used to train and test a predictive model in a lab environment. In ML workflows, the development data is typically split into training, validation, and hold-out test sets. The training set is used to learn the model parameters (e.g. the weights in a convolutional neural network), while the validation set is used during training to monitor the learning progress and avoid overfitting to the training set. The test set is (ideally) used only after training is completed, in order to quantify the performance of the model on 'unseen' data. It is prudent to avoid re-using the test data in development cycles as it can lead to unrealistic performance estimates (Dwork et al., 2015).

Importantly, the assumption that the performance of a trained model on the development test set is representative of the performance on new clinical data after deployment in varying environments is often violated due to differences in data characteristics, as discussed earlier. It is therefore absolutely critical to be able to clearly formalise and communicate the underlying assumptions regarding the data generating processes in the lab and real-world environments, which in turn can help anticipate and mitigate failure modes of the predictive system.

3.2.2 Causality in medical imaging

Given the specification of the input images, X, and the prediction targets, Y, it is imperative to determine which is the cause and which is the effect. Using the categorisation in Schölkopf et al. (2012), we wish to establish whether a task is

- *causal:* estimate $P(Y \mid X)$, when $X \to Y$ (predict effect from cause); or
- *anticausal:* estimate $P(Y \mid X)$, when $Y \to X$ (predict cause from effect).

The answer is crucial to all further causal analysis of the problem, and has a strong impact on the applicability of semi-supervised learning (Chapelle et al., 2006; Schölkopf et al., 2013)—discussed in more detail in Section 3.3.1—and on whether generative or discriminative models should be preferred (Blöbaum et al., 2015).

Recall the definitions of cause and effect: if the annotation could have been different by digitally editing the image beforehand, then one can conclude that the image causes the annotation. For example, manual segmentation masks are drawn over the image by visual inspection and would evidently be influenced by certain pixel changes. On the other hand, a pathology lab result would be unaffected by such manipulations. Images and targets may alternatively be confounded, i.e. descend



a Skin lesion classification

b Prostate tumour segmentation

Figure 3.2 Causal diagrams for medical imaging examples. Filled circular nodes represent measured variables, double circular nodes denote sample selection indicators, and squares are used for sample domain indicators. Here we additionally highlight the direction of the predictive task.

from a common cause. This relationship is often treated similarly to the anticausal case (Schölkopf et al., 2012).

It is generally possible to discern causal structures only when we are aware of the acquired data's background details, as meta-information plays a fundamental role in understanding the data generation and collection processes. A recently compiled ontology of medical imaging meta-information (Maier-Hein et al., 2018) contains several attributes that can help characterise the predictive causal direction in an imaging study, such as field of application and task category (e.g. lesion detection for screening, segmentation for treatment planning), as well as details about the annotation process (manual vs. (semi-)automatic vs. laboratory; image-wide vs. pixel-wise annotations; factors affecting reliability; etc.). Let us further illustrate this discussion with two practical examples, depicted in Fig. 3.2.

3.2.3 Worked clinical examples

Consider a skin lesion classification task, wherein a set of dermoscopic images (X) is collected along with histopathology diagnosis for melanoma following biopsy (Y). Here, Y is a gold-standard proxy for the true presence of skin cancer, and as such can be considered as a *cause* of the visual appearance of the lesion, X. This task is therefore *anticausal* (note the arrow directions in Fig. 3.2a). Further, routine dermoscopic examination of pigmented skin lesions typically results in a 'benign', 'suspicious', or 'malignant' label. Prediction of such labels would instead be *causal*, as they are obtained visually and could be affected if the images were digitally manipulated.

Now recall our earlier example where a team of radiologists had developed a new MRI-based diagnostic tool for prostate cancer. This time the team aims to improve the cancer treatment via radiotherapy by automating the planning process. Currently, the patient MRI scans (X) need to be manually segmented by carefully contouring the tumour regions and any organs-at-risk (Y). This annotation is done by visual inspection and evidently depends on image content, resolution, and contrast, for example, whereas manually editing the segmentation masks would have no effect on the images. These considerations allow us to conclude that image segmentation is a case of *causal*

prediction ($X \rightarrow Y$; see Fig. 3.2b).

For the two examples above, establishing the causal direction between images and prediction targets seemed reasonably straightforward. This is not always the case, and arguably in many settings identifying whether the relationship is causal or anticausal can be non-trivial, particularly if crucial meta-information is missing. Consider the case when prediction targets are extracted from radiology reports. At first, one may conclude that the report reflects purely the radiologist's reading of a medical image, hence image causes report. However, their conclusions could be based on additional information—potentially even more important than the findings in the images—such as blood tests or other diagnostic test results. In the context of segmentation, an annotator's knowledge about the grade of a tumour might influence how certain boundaries will be contoured, in which case an additional arrow from 'prostate cancer' to 'segmentation' could be included. This would however not alter the fact that the segmentations are a *consequence* of the images (and diagnoses), thus the task remains *causal*. Or what if image-derived diagnosis labels determined by an expert with long years of experience are nearly identical to biopsy results? Could these labels serve as proxies for the ground truth, configuring an anticausal relationship? These instances highlight the importance of investigating and modelling the full data generating process to make informed decisions about the causal relationships underlying the data. As there may not always be a single correct answer, it is crucial to clearly communicate the assumptions we make so these can be open to scrutiny.

3.3 Data scarcity

One of the notorious challenges in medical image analysis is the scarcity of labelled data, in great part due to the high costs of acquiring expert annotations or expensive lab tests, e.g. to confirm initial diagnosis. The techniques often used to circumvent this shortage, namely semi-supervised learning and data augmentation, have markedly different properties under the lens of causality.

3.3.1 Tackling data scarcity via semi-supervision

Semi-supervised learning (SSL) aims to leverage readily available unlabelled data in the hope of producing a better predictive model than is possible using only the scarce annotated data. Given this ambitious goal, it is perhaps unsurprising that strong requirements need to be met. Namely, the distribution of inputs needs to carry relevant information about the prediction task—otherwise it would be pointless to collect additional unlabelled data. This idea is typically articulated in terms of specific assumptions about the data which can be intuitively summarised as follows (Chapelle et al., 2006): similar inputs (images in our case) are likely to have similar labels and will naturally group into

clusters with high density in the input feature space. Lower density regions in that space in-between clusters are assumed to be ideal candidates for fitting decision boundaries of predictive models. In this context, considering large amounts of unlabelled data together with the scarce labelled data may reveal such low density regions and may lead to better decision boundaries than using labelled data alone.

Note how this idea insinuates an interplay between the distribution of inputs, $P(X)$, and the label conditional, $P(Y \mid X)$. Now recall that, by independence of cause and mechanism, if the prediction task is causal ($X \rightarrow Y$), then $P(X)$ is uninformative with respect to $P(Y \mid X)$, and SSL is theoretically futile in this case (Chapelle et al., 2006; Schölkopf et al., 2013). Since typical semantic segmentation tasks are *causal*, as illustrated in our prostate cancer example, there is likely little hope that semantic segmentation can fundamentally benefit from unlabelled data, which may relate to recent concerns raised in the literature (Oliver et al., 2018).

Intuitively, a model trained on image-derived annotations will attempt to replicate the (most often manual) annotation process, rather than to predict some pre-imaging ground truth (e.g. 'true' anatomy). It is plausible, then, that seeing more raw images without corresponding anatomical information provides no new insight about the annotation mechanism.

Conversely, if $Y \rightarrow X$ as for skin lesions, then these distributions *may* be dependent, and semi-supervision has a chance of success (Schölkopf et al., 2013). We conjecture that, in practice, anticausal problems are more likely than causal ones to comply with the SSL assumptions outlined above, as observed e.g. among the datasets analysed in Blöbaum et al. (2015).

That is not to say that SSL is completely useless for causal tasks, as there can be practical algorithmic benefits. Under certain conditions, unlabelled data can be shown to have a regularising effect, potentially boosting the accuracy of an imperfect model by lowering its variance (Cozman and I. Cohen, 2006), and may reduce the amount of labelled data required to achieve a given performance level (Balcan and A. Blum, 2006; Singh et al., 2008).

To the best of our knowledge, there have been no empirical studies systematically investigating the efficacy of SSL in causal and anticausal medical imaging tasks, especially for segmentation, hence further work is needed to validate its gains.

A recent comprehensive empirical study (Oliver et al., 2018) reported that properly tuned purely supervised models and models pre-trained on related labelled datasets (i.e. transfer learning) are often competitive with or outperform their semi-supervised counterparts. It also demonstrated that SSL can hurt classification performance under target shift (discussed as *prevalence shift* in Section 3.4.1) between labelled and unlabelled sets. This suggests that practitioners willing to apply SSL should be cautious of potential target distribution mismatch between labelled and unlabelled sets—e.g. unequal

proportions of cases and controls or presence of different pathologies.

3.3.2 Tackling data scarcity via data augmentation

In contrast with SSL, data augmentation produces additional plausible (x, y) pairs by systematically applying random, controlled perturbations to the data. Because it provides more information about the *joint* distribution $P(X, Y)$, rather than only the marginal $P(X)$, it is suitable for both causal and anticausal tasks, without the theoretical impediments of semi-supervised learning for causal prediction. This now ubiquitous technique is a powerful way of injecting domain knowledge to improve model robustness to variations one expects to find in the test environment. Notably, we can distinguish between augmentations encouraging *invariance* and *equivariance*.

Many tasks require predictions to be insensitive to certain types of variation. Examples include image intensity augmentations, such as histogram manipulations or addition of noise, and spatial augmentations (e.g. affine or elastic transformations) for image-level tasks (e.g. regression or classification, as in the skin lesion example). As these augmentations apply uniformly to all inputs x without changing the targets y, their benefits stem from a refined understanding of the conditional $P(X \mid Y)$, while contributing no new information about $P(Y)$.

For other tasks, such as segmentation or localisation, predictions must change similarly to the inputs, e.g. a spatial transformation applied to an image x—such as mirroring, affine or elastic deformations—should be likewise applied to the target y (e.g. spatial coordinates or segmentation masks, as in the prostate tumour example). Information is gained about the joint distribution via its shared spatial structure, related to e.g. anatomy and acquisition conditions.

3.4 Data mismatch

The recurrent issue of mismatch between data distributions, typically between training and test sets or development and deployment environments, tends to hurt the generalisability of learned models. In the generic case when no assumptions can be made about the nature of these differences, any form of learning from the training set is arguably pointless, as the test-time performance can be arbitrarily poor. Nonetheless, causal reasoning enables us to recognise special situations in which direct generalisation is possible, and to devise principled strategies to mitigate estimation biases. In particular, two distinct mechanisms of distributional mismatch can be identified: dataset shift and sample selection bias. Learning about their differences is helpful for diagnosing when such situations arise in practice.

3.4.1 Data mismatch due to dataset shift

Dataset shift is any situation in which the training and test data distributions disagree due to exogenous factors, e.g. dissimilar cohorts or inconsistent acquisition processes. As before, let X be the input images and Y be the prediction targets. We use an indicator variable D for whether we are considering the training ($P_{\text{tr}}(X, Y)$) or the test domain ($P_{\text{te}}(X, Y)$):[21]

$$P_{\text{tr}}(X, Y) := P_{D=0}(X, Y) \qquad \text{and} \qquad P_{\text{te}}(X, Y) := P_{D=1}(X, Y)\,. \tag{3.1}$$

For simplicity, in the following exposition we will refer only to disparities between training and test domains. This definition can however extend to differences between the development datasets (training and test data) and the target population (after deployment), when the latter is not well represented by the variability in the test data.

Moreover, when analysing dataset shift, it is helpful to conceptualise an additional variable Z, representing the unobserved physical reality of the subject's anatomy. We then interpret the acquired images X as imperfect and potentially domain-dependent measurements of Z, i.e. $Z \to X$.

Switching between domains may produce variations in the conditional relationships between X, Y, and Z or in some of their marginal distributions. Based on the predictive causal direction and on which factors of the joint distribution change or are invariant across domains, dataset shift can be classified into a variety of familiar configurations. Here we formulate the concepts of 'population shift', 'annotation shift', 'prevalence shift', 'manifestation shift', and 'acquisition shift'. These terms correspond roughly to particular dataset shift scenarios studied in general machine-learning literature, namely 'covariate shift', 'concept shift', 'target shift', 'conditional shift', and 'domain shift', respectively (Quiñonero-Candela et al., 2009). However, we believe it is beneficial to propose specific nomenclature that is more vividly suggestive of the phenomena encountered in medical imaging. By also explicitly accounting for the unobserved anatomy, the proposed characterisation is more specific and enables distinguishing cases that would otherwise be conflated, such as population or manifestation shift versus acquisition shift. The basic structures are summarised in Fig. 3.3 in the form of selection diagrams (causal diagrams augmented with domain indicators; Bareinboim and Pearl, 2016), and some examples are listed in Table 3.1. We hope this may empower researchers in our field to more clearly communicate dataset shift issues and to more easily assess the applicability of various solutions.

For causal prediction, we name *population shift* the case wherein only intrinsic characteristics (e.g. demographics) of the populations under study differ, i.e. $P_{\text{tr}}(Z) \neq P_{\text{te}}(Z)$. Fortunately, this case is *directly transportable*, i.e. a predictor estimated in one domain is equally valid in the other

[21] Here, we denote $P(\cdot \mid D) =: P_D(\cdot)$ to emphasise that the indicator D selects a different distribution in each domain.

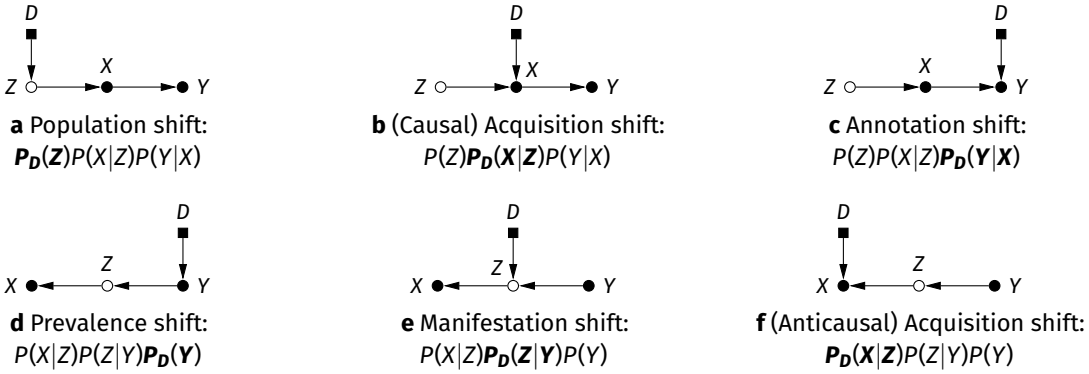


Figure 3.3 Selection diagrams for (a–c) causal and (d–f) anticausal dataset shift scenarios, with corresponding factorisations of the joint distribution $P_D(X,Y,Z)$. X is the acquired image; Y, the prediction target; Z, the unobserved true anatomy; and D, the domain indicator (0: 'train', 1: 'test'). An unfilled node means the variable is unmeasured.

Table 3.1 Types of dataset shift

Type	Direction	Change	Examples of differences
Population shift	causal	$P_D(Z)$	ages, sexes, diets, habits, ethnicities, genetics
Annotation shift	causal	$P_D(Y\|X)$	annotation policy, annotator experience
Prevalence shift	anticausal	$P_D(Y)$	baseline prevalence, case-control balance, target sel.
Manifestation shift	anticausal	$P_D(Z\|Y)$	anatomical manifestation of the target disease or trait
Acquisition shift	either	$P_D(X\|Z)$	scanner, resolution, contrast, modality, protocol

(Pearl and Bareinboim, 2014). An underfitted model ('too simple') may however introduce spurious dependencies, for which importance reweighting with $p_{\text{te}}(x)/p_{\text{tr}}(x)$ is a common mitigation strategy (Storkey, 2009; K. Zhang et al., 2015). This approach is not without limitations, however, as it requires access to $P_{\text{te}}(X)$ and may rely on further assumptions in order to truly correct for changes in $P(Z)$. Moreover, learning in this scenario makes sense only if the variability in the training data covers the support of the test distribution (Quiñonero-Candela et al., 2009)—in other words, there are no guarantees about extrapolation performance to modes of variation that are missing from the training environment.

Under *prevalence shift* (for anticausal tasks), the differences between datasets relate to class balance: $P_{\text{tr}}(Y) \neq P_{\text{te}}(Y)$. This can arise for example from different predispositions in the training and test populations, or from variations in environmental factors. If the test class distribution $P_{\text{te}}(Y)$ is known *a priori* (e.g. from an epidemiological study), generative models may reuse the estimated appearance model $P_{\text{tr}}(X \mid Y)$ $(= P_{\text{te}}(X \mid Y))$ in Bayes' rule, and, for discriminative models, instances can be weighted by $\frac{p_{\text{te}}(y)}{p_{\text{tr}}(y)}$ to correct the bias in estimating the training loss. Alternatively, more elaborate solutions based on the marginal $P_{\text{te}}(X)$ are possible (Storkey, 2009; K. Zhang et al., 2013), or the unknown target prevalence $P_{\text{te}}(Y)$ may be approximated using the confusion matrix of a trained

predictive model (Lipton et al., 2018).

Cases of *annotation shift* involve changes in class definitions, i.e. the same datum would tend to be labelled differently in each domain ($P_{\text{tr}}(Y \mid X) \neq P_{\text{te}}(Y \mid X)$). For example, it is not implausible that some health centres involved in an international project could be operating slightly distinct annotation policies or grading scales, or employing annotators with varying levels of expertise (e.g. senior radiologists vs. trainees). Without explicit assumptions on the mechanism behind such changes, models trained to predict $P_{\text{tr}}(Y \mid X)$ evidently cannot be expected to perform sensibly in the test environment, and no clear solution can be devised (Moreno-Torres et al., 2012). A tedious and time-consuming calibration of labels or (partial) re-annotation may be required to correct for annotation shift.

Another challenging scenario is that of *manifestation shift*, under which the way anticausal prediction targets (e.g. disease status) physically manifest in the anatomy changes between domains. In other words, $P_{\text{tr}}(Z \mid Y) \neq P_{\text{te}}(Z \mid Y)$. As with annotation shift, this cannot be corrected without strong parametric assumptions on the nature of these differences.

We lastly discuss *acquisition shift*, resulting from the use of different scanners or imaging protocols, which is one of the most notorious and well-studied sources of dataset shift in medical imaging (Glocker et al., 2019). Typical pipelines for alleviating this issue involve spatial alignment (normally via rigid registration and resampling to a common resolution) and intensity normalisation. In addition, the increasingly active research area of domain adaptation investigates data harmonisation by means of more complex transformations, such as extracting domain-invariant representations (Ganin et al., 2016; Kamnitsas et al., 2017a) or translating between imaging modalities (Frangi et al., 2018), e.g. synthesising MRI volumes from CT scans (Huo et al., 2019). Note that domain adaptation may fail or even be detrimental under changes in class prevalence (Arjovsky et al., 2019).

Returning to the prostate cancer example, suppose our dataset was collected and annotated for research purposes, employing a high-resolution 3 T MRI scanner and containing a majority of younger patients, and that the trained predictive model is to be deployed for clinical use with conventional 1.5 T scanners. This is a clear case of dataset shift, firstly because the images are expected to be of different quality (acquisition shift). Secondly, because the different age distribution in the target population entails variations in prostate size and appearance (population shift). In addition, the presence of both types of shift can lead to confounding (see Fig. 3.2b): a model trained on this data may erroneously learn that image quality is predictive of the risk of prostate cancer.



a Random **b** Image-dependent **c** Target-dependent **d** Jointly dependent

Figure 3.4 Causal diagrams for different sample selection scenarios. S=1 indicates an observed sample, and plain edges represent either direction.

Table 3.2 Types of sample selection

Type	Causation	Examples of selection processes	Resulting bias
Random	none	uniform random sampling	none
Image	$X \to S$	visual phenotype selection (e.g. anatomical traits, lesions)	population shift
		image quality control (QC; e.g. noise, low contrast, artefacts)	acquisition shift
Target	$Y \to S$	hospital admission, filtering by disease, annotation QC, learning strategies (e.g. class balancing, patch selection)	prevalence shift
Joint	$X \to S \leftarrow Y$	combination of the above (e.g. curated benchmark dataset)	selection bias (Hernán et al., 2004)

3.4.2 Data mismatch due to sample selection bias

A fundamentally different process that also results in systematic data mismatch is sample selection. It is defined as the scenario wherein the training and test cohorts come from the same population, though each training sample is measured ($S = 1$) or rejected ($S = 0$) according to some selection process S that may be subject-dependent:

$$P_{\text{tr}}(X, Y) := P(X, Y \mid S = 1) \quad \text{and} \quad P_{\text{te}}(X, Y) := P(X, Y)\,. \tag{3.2}$$

Some examples are presented in Table 3.2. The main difference to standard dataset shift is the *data-dependent* selection mechanism (Fig. 3.4), as opposed to external *causes* of distributional changes (Fig. 3.3). In other words, the indicator variables in sample selection concern alterations in the data-gathering process rather than in the data-generating process (K. Zhang et al., 2015).

Completely random selection simply corresponds to uniform subsampling, i.e. when the training data can be assumed to faithfully represent the target population ($P_{\text{tr}}(X, Y) \equiv P_{\text{te}}(X, Y)$). Since the analysis will incur no bias, the selection variable S can safely be ignored. We conjecture this will rarely be the case in practice, as preferential data collection is generally unavoidable without explicit safeguards and careful experimental design.

Selection can be affected by the appearance of each image in two different manners. We can select

subjects based on anatomical features—viewing the image X as a proxy for the anatomy Z—which has similar implications to population shift. Alternatively, selection criteria may relate to image quality (e.g. excluding scans with noise, poor contrast, or artefacts), which is akin to acquisition shift (Moreno-Torres et al., 2012). If selection is purely image-based ($X \to S$, cf. Fig. 3.4b), we may exploit the conditional independence $S \perp\!\!\!\perp Y \mid X$, which implies that the predictive relation is directly *recoverable* (Bareinboim et al., 2014), i.e. $P_{\text{te}}(Y \mid X) \equiv P_{\text{tr}}(Y \mid X)$. In a learning scenario, however, the objective function would still be biased, and methods for mitigating the corresponding cases of dataset shift can be employed. Relating back to the skin lesion example, patients are referred for biopsy only if dermoscopy raises suspicions. As inclusion in this study is image-dependent, a dataset with ground truth biopsy labels is not representative of the overall distribution of pigmented skin lesions.

When selection is solely target-dependent ($Y \to S$), we have $P_{\text{te}}(X \mid Y) \equiv P_{\text{tr}}(X \mid Y)$, and it can be treated as prevalence shift. This will typically result from factors like hospital admission, recruitment or selection criteria in clinical trials, or annotation quality control. Notably, machine-learning practitioners should be wary that it can also arise as a side-effect of certain training strategies, such as class re-balancing or image patch selection for segmentation (e.g. picking only patches containing lesion pixels).

Sample selection can additionally introduce spurious associations when the selection variable S is a common effect of X and Y (or of causes of X and Y): *implicitly conditioning* on S unblocks an undesired causal path between X and Y (see Methods). This is the classic situation called selection bias (Hernán et al., 2004; cf. Berkson's paradox, Pearl, 2009), and recovery is more difficult without assumptions on the exact selection mechanism. In general, it requires controlling for additional variables to eliminate the indirect influence of X on Y via conditioning on the collider S (Bareinboim and Pearl, 2016; Bareinboim et al., 2014).

3.5 Discussion

This chapter provides a fresh perspective on key challenges in machine learning for medical imaging using the powerful framework of causal reasoning. Not only do our causal considerations shed new light on the vital issues of data scarcity and data mismatch in a unifying approach, but the presented analysis can hopefully serve as a guide to develop new solutions. Perhaps surprisingly, causal theory also suggests that the common task of semantic segmentation may not fundamentally benefit from unannotated images via semi-supervision. This possibly controversial conclusion may prompt empirical research into validating the feasibility and practical limitations of this approach.

Other advanced topics could be worth exploring in future work for causally expressing more

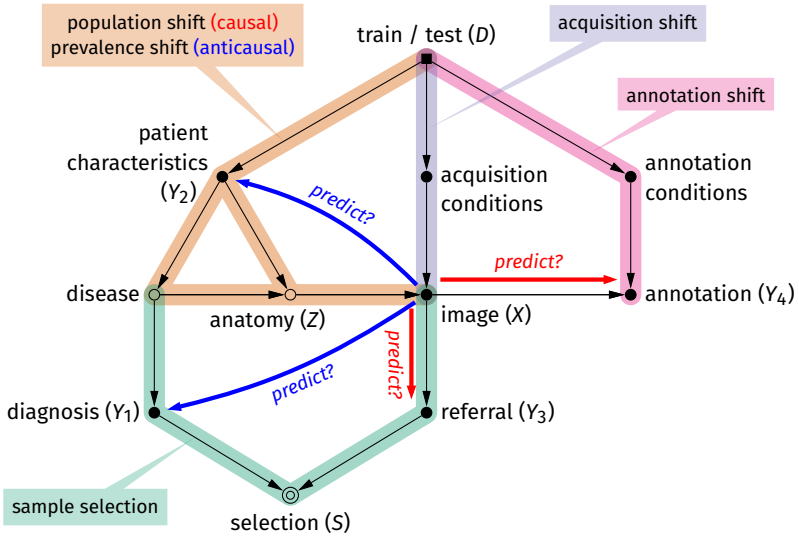


Figure 3.5 A 'scaffold' causal diagram summarising typical medical imaging workflows. Here are represented a variety of possible prediction targets (marked Y_1–Y_4): some anticausal (Y_1, Y_2) and others, causal (Y_3, Y_4). 'Annotation' here refers to any image-derived data, such as lesion descriptions, regions of interest, spatial landmark coordinates, or segmentation maps. Note that annotators will often be aware of the patients' records and diagnoses, in which cases there could be additional arrows from Y_1 or Y_2 towards Y_4.

subtle facets of predictive modelling workflows. In particular, one recurring topic in epidemiology and sociology that is relevant to our imaging context is *measurement bias* (Hernán and Cole, 2009; Shahar, 2009). This is the study of properties of proxy variables, which stand in for true variables of interest that are difficult or impossible to measure directly. Of particular note are the cases wherein proxies are additionally affected by other variables ('differential'), or when measurement errors for separate proxies are correlated ('dependent') (Lash et al., 2014). Measurement bias was explored here for the case of acquisition shift (images as proxies for anatomy, affected by the domain), and similar considerations could extend to other variables, e.g. patient records or pathology results.

A further pertinent topic is that of *missingness*. Whereas sample selection refers to the observability of full records, missingness concerns partial measurements—i.e. when some subjects may be missing observations of some variables. This is the context of semi-supervised learning, for example, as target labels are observed only for a subset of the data points. The classical characterisation distinguishes whether data is *missing completely at random*, *missing at random*, or *missing not at random*, when the missingness of a measurement is independent of any of the variables of interest, dependent on observed variables, or dependent on the missing values, respectively (Rubin, 1976). Causal diagrams again prove instrumental in identifying such structural assumptions about missingness mechanisms (Daniel et al., 2012; Mohan et al., 2013).

Finally, we highlight that our contribution is only the first step towards incorporating causality in medical image analysis. Here we introduce to this community purely the language of causal reasoning, hoping this will facilitate novel research directions exploiting causality theory to its full extent.

Table 3.3 Step-by-step recommendations

1. Gather meta-information about the data collection and annotation processes to reconstruct the full story of the dataset.
2. Establish the predictive causal direction: does the image cause the prediction target or vice versa? If annotations are scarce and image $\rightarrow$ target, semi-supervised learning may be futile, while data augmentation remains a viable alternative.
3. Identify any evidence of mismatch between datasets (Table 3.1). When applicable, importance reweighting is a common mitigation strategy; see further specific advice in the text.
   - If causal (image $\rightarrow$ target): population shift, annotation shift
   - If anticausal (target $\rightarrow$ image): prevalence shift, manifestation shift
4. Verify what types of differences in image acquisition are expected, if any. Consider applying data harmonisation techniques and domain adaptation (if test images are available).
5. Determine whether the data collection was biased with respect to the population of interest, and whether selection was based on the images, the targets, or both (Table 3.2). Refer to dataset shift guidance for mitigating the resulting biases.
6. Draw the full causal diagram including postulated direction, shifts, and selections.

Specifically, the endeavours of *causal inference* and *causal discovery* are so far largely unexplored in medical imaging. In this context, they could lead to the discovery of new imaging biomarkers and to exciting new applications such as personalised counterfactual predictions ('What if a patient were not a smoker?'). Large population imaging studies such as the UK Biobank (Conroy et al., 2019; Miller et al., 2016) can greatly empower this kind of research, as they offer unique opportunities for extracting the relevant patterns of variation from sheer observational data.

Beside enabling new research directions, incorporation of causal reasoning in medical image analysis aligns with a growing awareness among stakeholders of the need for responsible reporting in this field. There have been increasing efforts from regulatory bodies—such as the US Food and Drug Administration (Parikh et al., 2019; FDA, 2019), the UK's Department of Health and Social Care (UK DHSC, 2018), National Institute for Health and Care Excellence (NICE, 2019), and NHSX (Joshi and Morley, 2019), and even the World Health Organization (Wiegand et al., 2019)—to outline best practices for the safe development and monitoring of AI-enabled medical technologies ('Walking the tightrope of artificial intelligence guidelines in clinical practice' 2019). Guidelines for designing and reporting traditional clinical trials are now also being specialised for AI-based interventions (Liu et al., 2020; Rivera et al., 2020). This has been accompanied by a recent surge in discussion among the medical community about the opportunities and, crucially, the risks of deploying such tools in clinical practice (AOMRC, 2019; Ghassemi et al., 2019; Langlotz et al., 2019; Prevedello et al., 2019; N. H. Shah et al., 2019; N. D. Shah et al., 2018; Van Calster et al., 2019; Wiens et al., 2019). Most of the apprehension

revolves around the *external validity* of these predictive models, i.e. their generalisability beyond the development environment, in terms of e.g. robustness to dataset shift (Ghassemi et al., 2019; Prevedello et al., 2019) and selection bias (Ghassemi et al., 2019; S. Hoffman and Podgurski, 2013), as discussed herein. Other important concerns involve data inaccuracy, inconsistency, and availability (Ghassemi et al., 2019; S. Hoffman and Podgurski, 2013; Langlotz et al., 2019; Prevedello et al., 2019; N. D. Shah et al., 2018), and alignment of the model training objective with the target clinical setting (Prevedello et al., 2019; N. H. Shah et al., 2019; N. D. Shah et al., 2018; Wiens et al., 2019). In a similar yet complementary vein to the notable TRIPOD guidelines (Collins and Moons, 2019; Collins et al., 2015) and the recent SPIRIT-AI and CONSORT-AI extensions to traditional clinical trial checklists (Liu et al., 2020; Rivera et al., 2020), our work ties precisely into this context of encouraging transparent reporting of predictive analytics in healthcare.

This debate also relates to parallel initiatives from within the machine learning community, in specific in the emerging field of fairness, accountability, and transparency (FAT). Scholars in FAT have proposed checklist-style guidelines for reporting datasets (Gebru et al., 2018) and models (Mitchell et al., 2019), for example, and have been investigating sources of failure for machine-learning models, among which is poor reporting (Saria and Subbaswamy, 2019). Interestingly, the same formalism of causal reasoning explored here was also shown to be especially well-suited for expressing and addressing issues of unfairness (e.g. societal biases; Chiappa and Isaac, 2019; Kilbertus et al., 2017; Kusner et al., 2017) and dataset shift (Subbaswamy et al., 2019) in other contexts.

Overall, the goal of this work has been to introduce to the medical imaging community the language of causal diagrams, and to demonstrate how it can illuminate common issues in predictive modelling. While causal reasoning by itself may not solve any of the data scarcity or mismatch problems, it provides a clear and precise framework for expressing assumptions about the data. Presenting such assumptions transparently in the form of causal diagrams makes them immediately recognisable by other researchers, and therefore easier to be confirmed or disputed. The real challenge lies in identifying these very assumptions, as they can often be unclear or ambiguous.

To facilitate this task, we offer in Table 3.3 a step-by-step summary of our recommendations, and Fig. 3.5 presents a generic 'scaffold' diagram from which most typical workflows can be adapted. Readers may then refer to the other tables for help in identifying the components of their own diagram for the problem at hand. We believe that this exercise of building the full causal story of a dataset will encourage analysts to consider potential underlying biases more thoroughly, and that it may, like the TRIPOD checklist, lead to 'more comprehensive understanding, conduct, and analysis of prediction model studies' (Collins and Moons, 2019).

So far, we have analysed high-level considerations obtained via causal reasoning. In the remainder

of this thesis, we will focus on more practical aspects of probabilistic modelling, first presenting alternative approaches based on nonparametrics, then a testbed for learning with heterogeneous data, and ultimately a framework combining causal structures and deep generative models. Specifically, the next chapter explores a probabilistic model for image intensities, which is later used in a novel method for addressing a form of acquisition shift (Section 3.4.1).

Chapter 4

Modelling and normalising image intensities

This chapter contains material from my Master's thesis:

- **Daniel C. Castro** (2016). 'Dirichlet process mixture models applied to brain image segmentation'. MRes thesis. London: Imperial College London

and is based on the following publications:

- **Daniel C. Castro** and Ben Glocker (2016). 'Dirichlet process mixture models: application to brain image segmentation'. In: *NIPS 2016 Workshop on Practical Bayesian Nonparametrics* (Barcelona, Spain). HDL: 10044/1/65183;
- **Daniel C. Castro** and Ben Glocker (2018). 'Nonparametric density flows for MRI intensity normalisation'. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018* (Granada, Spain). Vol. 11070. LNCS. Springer, Cham, pp. 206–214. DOI: 10.1007/978-3-030-00928-1_24. arXiv: 1806.02613.

Code for the intensity normalisation tool is available at https://github.com/dccastro/NDFlow.

This chapter begins with an investigation of the feasibility of nonparametric mixtures for modelling multimodal image intensities. The flexibility of nonparametric models makes them particularly suitable for medical imaging applications, where intensity distributions can show extreme variability and it is hard to make prior assumptions on the correct model structure. We validate the use of multivariate Dirichlet process Gaussian mixture models (DPGMM) with two unsupervised brain MRI segmentation tasks (normal tissues and traumatic injury lesions). The preliminary segmentation results showed great promise, and the DPGMMs demonstrated their capability to learn intricate multidimensional probability distributions.

The remainder of the chapter expands the application scope of these models, presenting a novel approach to image intensity normalisation. Alongside spatial registration, normalisation is an essential tool for alleviating the pervasive issue of acquisition shift in imaging (Section 3.4.1). The proposed

scheme is based on matching histogram densities modelled as DPGMMs. The source mixture model is transformed to minimise its L^2 divergence towards a target model, then the voxel intensities are transported through a mass-conserving flow to maintain agreement with the moving density. Our multi-centre study with brain MRI data shows that the proposed technique matches histograms accurately and smoothly, while substantially improving the consistency of tissue intensities over a baseline affine transformation. Finally, we verify that nonlinear intensity normalisation is a step towards effective imaging data harmonisation.

4.1 Unsupervised segmentation

In medical image segmentation, common approaches are based on supervised learning (e.g. Kamnitsas et al., 2017b), atlas-based methods (Cabezas et al., 2011), unsupervised techniques, such as mixture models (e.g. Ashburner and Friston, 2005), or combinations thereof (e.g. Ledig et al., 2015, 2012).

Unsupervised methods are attractive because labelled medical imaging data is relatively scarce (Section 3.3), and they essentially attempt to discover structure in the data without any explicit constraints on the output mapping. Furthermore, nonparametric methods make no hard assumptions about this structure, dynamically adapting their complexity as more data is observed (Ghahramani, 2013). In particular, Dirichlet process mixture models (DPMMs; Antoniak, 1974; Ferguson, 1983) can be seen as an infinite extension of finite mixture models. They place no bound on the number of mixture components, and can, in principle, fit arbitrarily complex distributions. This property makes DPMMs very attractive for modelling elaborate data such as multimodal MRI.

DPMMs have recently been used in a variety of applications in medical imaging. Thirion et al. (2007) and S. Kim and Smyth (2006) have used DPMMs for modelling spatial activation patterns in functional imaging. Jbabdi et al. (2009) proposed using hierarchical Dirichlet processes for modelling intra- and inter-subject variability in connectivity-based parcellation and Wang et al. (2011) used DPMMs for tractography segmentation. In the context of MRI segmentation, Wachinger and Golland (2014) presented a formulation in which DPMMs were used to model background variation when segmenting small structures to reduce bias. Ferreira da Silva (2007, 2009) performed whole-brain MRI segmentation with a univariate DPMM. More generally, DP-based models have also been applied in computer vision for natural image segmentation (e.g. Sudderth and Jordan, 2009; Torralba et al., 2006).

In practice, DPMMs are predominantly applied with Gaussian components, referred to as a DP Gaussian mixture model (DPGMM). Ferreira da Silva's work was restricted to a single MRI modality (T1), with all its univariate DPGMM components constrained to share the same variance. In this work,

we have addressed *multimodal* brain image segmentation with a multivariate DPGMM, allowing full and untied covariances for all mixture components. Because each imaging modality carries complementary information about the underlying object (in this case, a human brain), we expect that a model capable of leveraging this wealth of information could generate more accurate segmentations.

4.2 Nonparametric intensity model

In this work, we have studied a simple voxel-wise intensity model. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ represent the N voxels of an image, where $\mathbf{x}_i \in \mathbb{R}^D$ are the intensities of voxel i across the D available modalities (assumed perfectly registered).

4.2.1 Intensity model

The generative process for the Dirichlet process Gaussian mixture model (DPGMM) used in this work is formulated as follows:[22]

$$\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha) \tag{4.1a}$$

$$\boldsymbol{\Lambda}_k \sim \mathcal{W}(a, \mathbf{B}), \qquad k = 1, 2, \ldots \tag{4.1b}$$

$$\boldsymbol{\mu}_k \,|\, \boldsymbol{\Lambda}_k \sim \mathcal{N}\big(\boldsymbol{\mu}_0, (\nu \boldsymbol{\Lambda}_k)^{-1}\big), \qquad k = 1, 2, \ldots \tag{4.1c}$$

$$z_i \,|\, \boldsymbol{\pi} \sim \mathrm{Cat}(\boldsymbol{\pi}), \qquad i = 1, \ldots, N \tag{4.1d}$$

$$\mathbf{x}_i \,|\, z_i, (\boldsymbol{\mu}_k)_k, (\boldsymbol{\Lambda}_k)_k \sim \mathcal{N}\big(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Lambda}_{z_i}^{-1}\big), \qquad i = 1, \ldots, N, \tag{4.1e}$$

where $\mathrm{GEM}(\alpha)$ is the stick-breaking distribution (Section 2.2.1.1).

4.2.2 Inference

In both our experiments, we have set hyperparameters $\nu = 1$ and $a = D + 1$, while $\boldsymbol{\mu}_0$ and $\mathbf{B}$ were set to match the first and second moments of the prior predictive distribution, $p(\mathbf{x})$, to the data. Inference was done in the truncated DPMM variational framework of Blei and Jordan (2006), using a truncation level K. Here, we shall approximate the full posterior distribution over model parameters $\{\boldsymbol{\beta}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\Lambda}_{1:K}\}$ (note we use the stick-breaking variables $\boldsymbol{\beta}$ for $\boldsymbol{\pi}$) and local latent variables $\mathbf{Z}$ by

$$p(\boldsymbol{\beta}, \boldsymbol{\mu}_{1:K}, \boldsymbol{\Lambda}_{1:K}, \mathbf{Z} \,|\, \mathbf{X}) \approx \prod_{k=1}^{K-1} q(\beta_k) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k) \prod_{k=1}^{K} q(\boldsymbol{\Lambda}_k) \prod_{n=1}^{N} q(z_n)\,. \tag{4.2}$$

[22] We consider the parametrisation of the Wishart distribution $\mathcal{W}$ given by Bernardo and A. F. M. Smith (2000), assuming $\mathbf{X} \in \mathbb{R}^{D \times D}$ is symmetric positive-definite, $a > \frac{D-1}{2}$, and $\mathbf{B}$ is symmetric non-singular: $\mathcal{W}_D(\mathbf{X} \,|\, a, \mathbf{B}) = (\Gamma_D(a))^{-1} |\mathbf{B}|^a |\mathbf{X}|^{a - \frac{D+1}{2}} \exp(-\mathrm{tr}[\mathbf{B}\mathbf{X}])$, where $\Gamma_D(\cdot)$ is the multivariate gamma function.

Table 4.1 True posterior and variational distributions for the multivariate DPGMM. The primed variables are the parameters of the corresponding posterior distributions given the observations assigned to cluster k.

Posterior conditionals	Variational
$p(\boldsymbol{\Lambda}_k \mid \mathbf{X}, \mathbf{Z}) = \mathcal{W}(\boldsymbol{\Lambda}_k \mid a'_k, \mathbf{B}'_k)$	$q(\boldsymbol{\Lambda}_k) = \mathcal{W}(\boldsymbol{\Lambda}_k \mid c_k, \mathbf{D}_k)$
$p(\boldsymbol{\mu}_k \mid \mathbf{X}, \mathbf{Z}, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k \mid \boldsymbol{\mu}_0{}'_k, (\nu'_k \boldsymbol{\Lambda}_k)^{-1})$	$q(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_k \mid \mathbf{m}_k, \boldsymbol{\Psi}_k)$
$p(\beta_k \mid \mathbf{Z}) = \mathcal{B}(\beta_k \mid 1 + N_k, \alpha + N_{>k})$	$q(\beta_k) = \mathcal{B}(\beta_k \mid \gamma_{k,1}, \gamma_{k,2})$
$p(z_n = k \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}) \propto \pi_k \mathcal{N}_D(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})$	$q(z_n = k) = \phi_{nk}$

Table 4.1 summarises the variational distributions employed for each variable, side-by-side with the true posteriors for comparison (note that these are conditional on all other variables, whereas we wish to approximate the full joint distribution). Below we list the derived mean-field parameter updates (see Section 2.2.2.1) for the DPGMM, letting $\widehat{N}_k = \sum_n \phi_{nk}$ and $\widehat{N}_{>k} = \sum_{l>k} \widehat{N}_l$.

- Mean parameters: $q(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_k \mid \mathbf{m}_k, \boldsymbol{\Psi}_k)$

$$\mathbf{m}_k = \frac{\nu \boldsymbol{\mu}_0 + \sum_n \phi_{nk} \mathbf{x}_n}{\nu + \widehat{N}_k} \qquad \boldsymbol{\Psi}_k = \frac{1}{(\nu + \widehat{N}_k) c_k} \mathbf{D}_k \tag{4.3a}$$

- Precision parameters: $q(\boldsymbol{\Lambda}_k) = \mathcal{W}(\boldsymbol{\Lambda}_k \mid c_k, \mathbf{D}_k)$

$$\begin{aligned} c_k &= a + \frac{\widehat{N}_k + 1}{2} \\ \mathbf{D}_k &= \mathbf{B} + \frac{1}{2}\left[\sum_n \phi_{nk} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^\mathsf{T} + \nu (\mathbf{m}_k - \boldsymbol{\mu}_0)(\mathbf{m}_k - \boldsymbol{\mu}_0)^\mathsf{T} + (\nu + \widehat{N}_k) \boldsymbol{\Psi}_k\right] \end{aligned} \tag{4.3b}$$

- Stick-breaking parameters: $q(\beta_k) = \mathcal{B}(\beta_k \mid \gamma_{k,1}, \gamma_{k,2})$

$$\gamma_{k,1} = 1 + \widehat{N}_k \qquad \gamma_{k,2} = \alpha + \widehat{N}_{>k} \tag{4.3c}$$

- Assignment parameters: $q(z_n = k) = \phi_{nk}$

$$\begin{aligned} \log \phi_{nk} = \frac{1}{2}\Big(\psi_D(c_k) - \log|\mathbf{D}_k| - c_k \operatorname{tr}\Big[\mathbf{D}_k^{-1}\big((\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^\mathsf{T} + \boldsymbol{\Psi}_k\big)\Big]\Big) \\ + [\psi(\gamma_{k,1}) - \psi(\gamma_{k,1} + \gamma_{k,2})] + \sum_{l<k} [\psi(\gamma_{l,2}) - \psi(\gamma_{l,1} + \gamma_{l,2})] + \text{const.}, \end{aligned} \tag{4.3d}$$

where $\psi(x) = \frac{\mathrm{d}}{\mathrm{d}x} \log \Gamma(x)$ and $\psi_D(x) = \sum_{d=1}^{D} \psi\left(x + \frac{1-d}{2}\right)$ are the digamma and multivariate digamma functions, respectively.

4.3 Segmentation experiments

4.3.1 Evaluation

To assess the compatibility of the cluster partitioning with the ground-truth labels, we have computed a number of clustering consistency metrics: adjusted Rand index (ARI; Hubert and Arabie, 1985; Rand, 1971), normalised mutual information (NMI), v-measure (homogeneity, completeness and validity; Rosenberg and Hirschberg, 2007), and purity.

We are mostly concerned with correctly delineating the structures of interest, regardless of inner subdivisions. To compute spatial accuracy metrics, which require matching labels, we have merged clusters by reassigning their labels to the majority ground-truth label in each cluster. The chosen spatial metrics were the Dice similarity coefficient (DSC; Dice, 1945), the average symmetric surface distance (ASSD), the Hausdorff distance (HD), and the relative absolute volume difference (RAVD).

4.3.2 Normal tissue segmentation

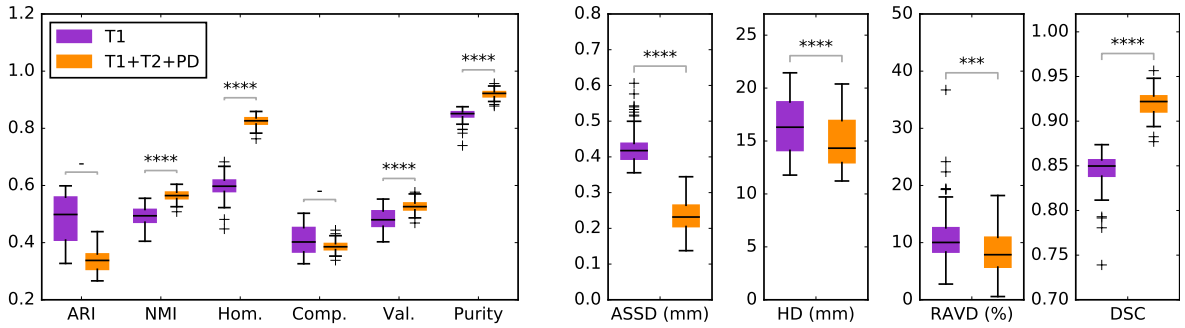
In our first experiment, we aimed to assess whether the Dirichlet process Gaussian mixture model could successfully segment grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF) in multimodal images of healthy brains, using a subset of the IXI database.[23] It comprises registered T1- and T2-weighted and proton density (PD) MRI scans of 86 healthy subjects aged 35.8 ± 11.9, with dimensions $256 \times 256 \times 150$ and $.9375 \times .9375 \times 1.2\,\mathrm{mm}^3$ resolution. Since there are no manual segmentations available in this dataset, reference GM, WM, and CSF probability maps were computed with SPM12[24] to obtain 'gold standard' segmentations.

As a baseline, we fitted a univariate DPGMM to the single T1 image volumes, as was done in the work of Ferreira da Silva (2007, 2009). We then trained multivariate DPGMMs on the combined T1, T2, and PD volumes and compared the results.

The consistency results are plotted in Fig. 4.1a. The model found 11.9 ± 0.3 components when using multimodal data and 8.1 ± 1.1 components with a single modality—even though all runs had the same fixed concentration, $\alpha = 0.5$ and truncation level 20. This more fragmented segmentation would explain the observed drop in completeness and in Rand index, because pairs of voxels belonging to the same tissue are more likely to be split across cluster borders.

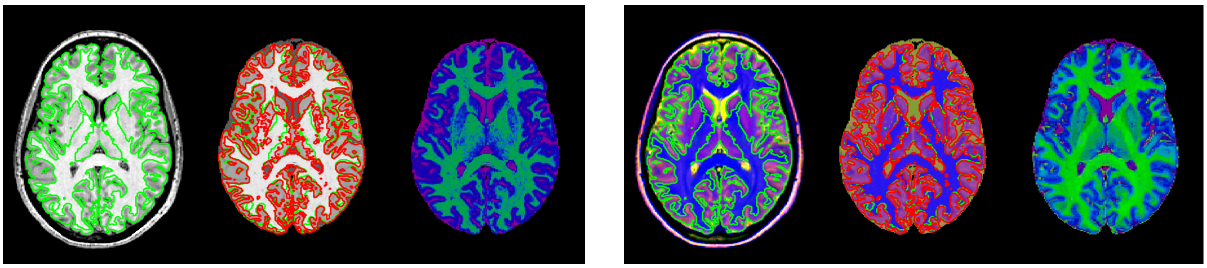
On the other hand, the multimodal clusters are substantially purer and more homogeneous—which pulls up the validity, counteracting the slight decrease in completeness—and convey more information about the reference tissue labels. These effects indicate that the availability of the

[23] http://brain-development.org/ixi-dataset/
[24] http://www.fil.ion.ucl.ac.uk/spm/software/spm12/



a Clustering consistency metrics **b** Spatial accuracy metrics

Figure 4.1 Results for the normal tissue segmentation experiment. The boxplots indicate quartiles, median, range and outliers. One to four stars indicate $p < 0.05, 0.01, 10^{-3}$, and 10^{-4}, respectively, and '–' is insignificant (one-sided Wilcoxon signed-rank test).



a T1 slice. K = 7 components, ASSD = 0.49 mm, HD = 16.0 mm, RAVD = 5.47 %, DSC = 0.822.

b PD, T2, and T1 slice (RGB). K = 12 components, ASSD = 0.30 mm, HD = 16.0 mm, RAVD = 3.99 %, DSC = 0.906.

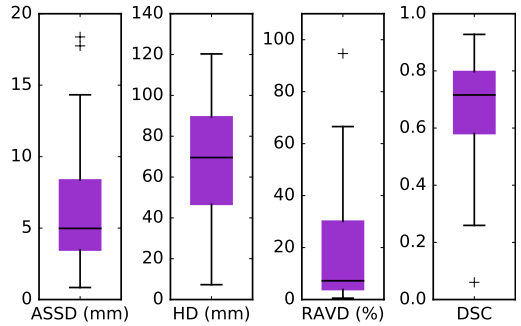
Figure 4.2 Uni- and multimodal grey matter (GM) segmentations of IXI subject 52. *Left:* Original image and overlaid reference segmentation (green). *Middle:* Fuzzy image reconstruction from cluster means, overlaid with reference (green) and obtained merged segmentation (red). *Right:* Cluster label map, weighted by the posterior label probabilities.

additional modalities allowed the model to infer more specialised clusters.

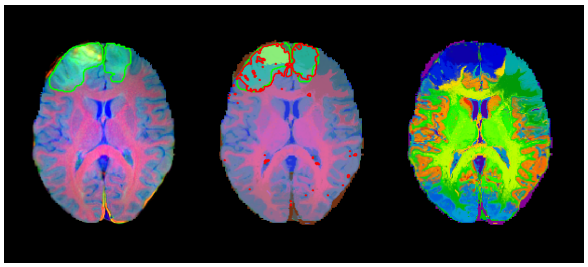
Regarding spatial accuracy (Fig. 4.1b), the multimodal model provided dramatically better matches to the reference segmentation, as indicated by the average surface distance and Dice coefficient. For each subject, the boundaries of the multimodal clusters follow the tissue boundaries 44 % more closely and present 8.9 % higher overlap than in the single modality case, on average. The subtle decrease in Hausdorff distance and volume difference suggests slightly fewer gross mistakes.

4.3.3 Lesion discovery

This experiment was intended as a proof of concept—rather than a thorough assessment—of the applicability of DPGMMs to the task of lesion segmentation. As such, it was run on a single axial slice for each subject instead of using the full volumes, which would have taken considerably longer to compute due to the higher dimensionality (five input modalities) and much greater number of components to estimate ($\alpha = 5$ and truncation level 40). This relatively high concentration was chosen for increased sensitivity to subtle intensity variations and potentially very small clusters. The selected



a Spatial accuracy metrics



b TBI subject 17744. MPRAGE, FLAIR, and GE overlaid as RGB. K = 31 components, ASSD = 4.9 mm, HD = 118 mm, RAVD = 4.7 %, DSC = 0.854.

Figure 4.3 Lesion segmentation results and example. *Left:* Original image and reference segmentation. *Middle:* Image reconstruction, with reference and obtained segmentations. *Right:* Cluster label map.

slices were the ones containing the most lesion voxels.

We analysed images of a cohort of 63 moderate-to-severe traumatic brain injury (TBI) patients. The dataset consists of MPRAGE, FLAIR, GE, T2, and PD scans, plus manual lesion segmentations. Each modality was skull-stripped, affinely registered to MNI space and resampled to 193 × 229 × 193 isotropic 1 mm^3 resolution. This data was originally collected and prepared for Kamnitsas et al. (2017b), where further details can be found.

For this task, because there is only one label, we computed only spatial accuracy metrics, plotted in Fig. 4.3a. Note that the simple majority merging procedure described in Section 4.3.1 failed to match any lesion for 17 of the 63 subjects, which have been excluded from the results. Interestingly, these subjects were among those with the smallest lesions.

The annotated lesions were correctly localised in most cases, and the surface distances and volume differences are remarkably low, given the difficulty of the task. Figure 4.3b shows the model is able to capture some of the finer inner structure of lesions. Further, we observe that the model has allocated additional clusters to accommodate bias field inhomogeneities. Because manual annotations involve substantial expert knowledge, our intensity-based model occasionally clustered together regions that had similar intensity profiles but were not both annotated as lesions, resulting in excessive false positives or false negatives.

4.3.4 Model visualisation

Finally—and most relevantly to the next part of this chapter—we analysed the density estimation capabilities of the DPGMM. In other words, we can evaluate the mixture density by marginalising over the cluster indicators z_i. As exemplified in Fig. 4.4 for an arbitrary subject in the dataset, we found that the DPGMM is able to faithfully capture not just one-dimensional intensity distributions, but also the full joint distributions of intensities across multiple imaging modalities.

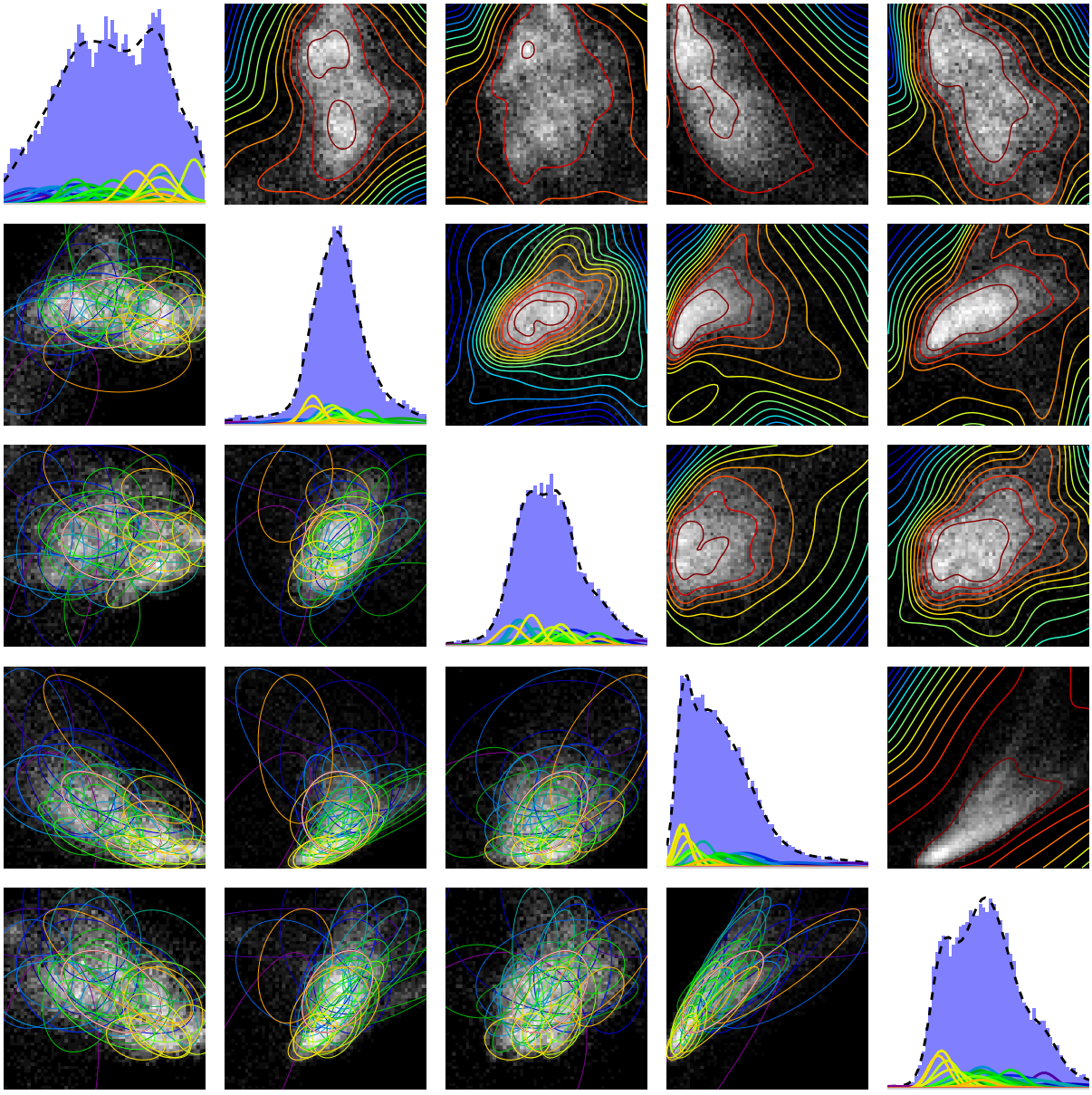


Figure 4.4 Empirical and fitted five-dimensional distributions for TBI subject 17770. The rows (top to bottom) and columns (left to right) correspond to MPRAGE, FLAIR, GE, T2, and PD modalities. Plots on the same column share their x-axis, and off-diagonal plots on the same row share the y-axis. Best viewed in colour.

Histograms: On the diagonal are the intensity histograms of each individual modality (in light blue). Off-diagonal, we have the two-dimensional log-histograms for each pair of modalities (column modality on x-axis vs. row modality on y-axis, lighter is higher).

Mixture components: The 1D and 2D marginals of the mixture components ($p(x^d \mid z = k)$ and $p(x^{d_1}, x^{d_2} \mid z = k)$) are overlaid on the histograms, rendered as bell curves on the diagonal and as ellipses on the lower left half, with matching colours. The mixture proportions (π_k) are represented by the scaling factor of the bell curves and thickness of the ellipse contours.

Likelihoods: The dashed line on the diagonal shows the 1D marginal likelihood ($p(x^d)$). On the upper right half, we see level curves of the 2D marginal log-likelihoods ($\log p(x^{d_1}, x^{d_2})$). Note the remarkably tight fit to the histograms in all 1D and 2D projections.

4.4 Intensity normalisation

Many medical image analysis methods rely on the hypothesis that corresponding anatomical structures present similar intensity profiles. Unlike computed tomography, magnetic resonance imaging does not produce scans in an absolute standard scale, in general. Even when using the same imaging protocols, there can be significant variation between different scanners. Acquisition parameters have a complex effect on the luminance of the acquired images, therefore a simple linear rescaling of intensities is usually insufficient for effective data harmonisation (Hellier, 2003). Therefore, a crucial factor for enabling the construction of large-scale image databases from multiple sites is accurate nonlinear intensity normalisation.

A number of different approaches have been introduced for this task (cf. Bergeest and Jäger, 2008), the most widely adopted of which is that of Nyúl et al. (2000), extending the original approach of Nyúl and Udupa (1999).[25] The authors proposed to normalise intensities by matching a set of histogram quantiles, using these as landmarks for a piecewise linear transformation. Despite its apparent simplicity, it has proven very effective in clinical applications (M. Shah et al., 2011).

Our proposed method, nonparametric density flows (NDFLOW), is perhaps conceptually closest to Hellier (2003)'s approach, which involves matching Gaussian mixture models (GMMs) fitted to a pair of image histograms. The author used a finite mixture to represent a pre-defined set of five tissues classes, whereas we propose to use nonparametric mixtures, focusing on accurately modelling the density rather than discriminating tissue types, and sidestepping the problem of pre-selecting the number of components. A further difference is that, instead of polynomially interpolating between the means of corresponding components, we build a smooth transformation based on density flows.

4.5 Nonparametric density matching

We begin by justifying and describing the density model used to represent the intensity distributions to be matched. We then introduce the chosen objective function with its gradients for optimisation. Finally, we present our flow-based transformation model, which deforms the data so it conforms to the matched density model. Note that we focus here on single-modality intensity normalisation, although the entire formulation below extends naturally to the multivariate case.

[25] For example, it is the algorithm used for histogram matching in the popular Insight Toolkit (ITK) software package.

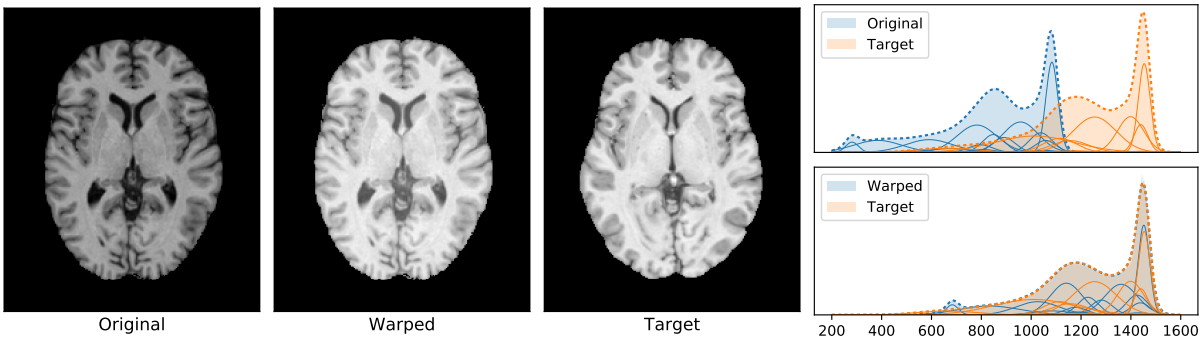


Figure 4.5 Comparison of two MRI scans, before and after the proposed NDFLOW normalisation. *Right:* histograms (shaded) and fitted mixture models (*dotted:* likelihood, *solid:* mixture components).

4.5.1 Intensity model

In order to be able to match the intensity distributions of a pair of images, a suitable probability density model is required. Typically, finite mixture models are considered for this task (Hellier, 2003; Roy et al., 2007). However, a well-known limitation of these is the requirement to specify *a priori* a fixed number of components, which may in addition call for an iterative model selection loop (e.g. Roy et al., 2007).

On the opposite end of the spectrum, another approach is to use kernel density estimation, which is widespread for shape registration (e.g. Hasanbelliu et al., 2011; Jian and Vemuri, 2011). However, this formulation would result in an unwieldy optimisation problem, involving thousands or millions of parameters and all pairwise interactions. Furthermore, the derived transformation would likely not be satisfactorily smooth without additional regularisation.

To overcome both issues we propose to once again use DPGMMs. Instead of specifying a fixed number of components, they rely on a vague concentration parameter, which regulates the expected amount of clustering fragmentation and enables them to adapt their complexity to the data at hand. By allowing an unbounded number of components and setting a versatile prior on the mixture proportions, they appear as a parsimonious middle ground for flexibility and tractability. As in the previous segmentation experiments, we fit the DPGMMs to each image's intensities using variational inference (Section 4.2.2).

4.5.2 Density matching

The first step is to perform a coarse affine alignment by matching the moving density's first and second moments to the target's, accounting for arbitrary translation and rescaling of the values. This same affine transformation is then also applied to the data before the nonlinear warping takes place.

We quantify the disagreement between two probability density functions q and p on a probability

space $\mathcal{X}$ by means of the L^2 divergence:

$$D_{L^2}[q,p] = \tfrac{1}{2}\|q-p\|^2 = \tfrac{1}{2}\|q\|^2 + \tfrac{1}{2}\|p\|^2 - \langle q,p\rangle\,, \tag{4.4}$$

where $\langle q,p\rangle = \int q(x)\,p(x)\,\mathrm{d}x$ is the L^2 inner product and $\|q\| = \sqrt{\langle q,q\rangle}$ is its induced norm. Aside from being symmetric, this quantity is positive and reaches zero iff $q \overset{\text{a.e.}}{=} p$. Crucially, unlike the usual Kullback–Leibler divergence, it is expressible in closed form for Gaussian mixture densities.[26]

Recall the following easily demonstrated properties of Gaussian densities:

Lemma 4.1. *Let $f(x) = \mathcal{N}(x\,|\,\mu,\lambda^{-1})$ and $g(x) = \mathcal{N}(x\,|\,\nu,\omega^{-1})$ be two Gaussian probability density functions. Their L^2 inner product and its derivatives w.r.t. the parameters of f are given by*

$$\begin{aligned}
\langle f,g\rangle &= \mathcal{N}(\mu\,|\,\nu,\lambda^{-1}+\omega^{-1})\,,\\
\partial_\mu\langle f,g\rangle &= \langle f,g\rangle\cdot\frac{\mu-\nu}{\lambda^{-1}+\omega^{-1}}\,,\\
\partial_\lambda\langle f,g\rangle &= \langle f,g\rangle\cdot\frac{1}{2}\left[\lambda^{-1}-(\lambda+\omega)^{-1}-\left(\omega\frac{\mu-\nu}{\lambda+\omega}\right)^2\right].
\end{aligned}$$

Let $q = \sum_k \pi_k q_k$ and $p = \sum_m \tau_m p_m$ denote two Gaussian mixtures, with components $q_k(x) = \mathcal{N}(x\,|\,\mu_k,\lambda_k^{-1})$ and $p_m(x) = \mathcal{N}(x\,|\,\nu_m,\omega_m^{-1})$. Equation (4.4) has tractable derivatives w.r.t. the parameters of q, which we use to optimise its components' means $(\mu_k)_k$ and precisions $(\lambda_k)_k$. Using the results in Lemma 4.1, Eq. (4.4) can be written explicitly as

$$\begin{aligned}
D_{L^2}[q,p] = \frac{1}{2}\sum_{l,k}\pi_l\pi_k\mathcal{N}(\mu_l\,|\,\mu_k,\lambda_l^{-1}+\lambda_k^{-1}) + \frac{1}{2}\sum_{m,n}\tau_m\tau_n\mathcal{N}(\nu_m\,|\,\nu_n,\omega_m^{-1}+\omega_n^{-1})\\
-\sum_{m,k}\tau_m\pi_k\mathcal{N}(\nu_m\,|\,\mu_k,\omega_m^{-1}+\lambda_k^{-1})\,.
\end{aligned}$$

Its gradients are also simply expressed as

$$\frac{\partial D_{L^2}[q,p]}{\partial\mu_k} = \sum_l w_{lk}\frac{\mu_l-\mu_k}{\lambda_l^{-1}+\lambda_k^{-1}} - \sum_m v_{mk}\frac{\nu_m-\mu_k}{\omega_m^{-1}+\lambda_k^{-1}}\,, \tag{4.5}$$

$$\begin{aligned}
\frac{\partial D_{L^2}[q,p]}{\partial\lambda_k} = \sum_l\frac{w_{lk}}{2}\left[\lambda_k^{-1}-(\lambda_l+\lambda_k)^{-1}-\left(\lambda_l\frac{\mu_l-\mu_k}{\lambda_l+\lambda_k}\right)^2\right]\\
-\sum_m\frac{v_{mk}}{2}\left[\lambda_k^{-1}-(\omega_m+\lambda_k)^{-1}-\left(\omega_m\frac{\nu_m-\mu_k}{\omega_m+\lambda_k}\right)^2\right],
\end{aligned} \tag{4.6}$$

where $w_{lk} = \pi_l\pi_k\langle q_l,q_k\rangle$ and $v_{mk} = \tau_m\pi_k\langle p_m,q_k\rangle$. To make sure the precisions $\{\lambda_k\}_k$ remain non-negative throughout the optimisation, we can simply reparametrise them as $\lambda_k = \ell_k^2$, with $\ell_k\in\mathbb{R}$.

We have found, in practice, that it is largely unnecessary to adapt the mixing proportions, $(\pi_k)_k$, to

[26] Although we did some preliminary investigation using the related Cauchy–Schwarz divergence (Principe et al., 2000), minimisation of the L^2 divergence seems to have more well-behaved gradients and to be much more stable for this application.

get an excellent agreement between mixture densities. In fact, changing the mixture weights would require transferring samples *between* mixture components. Although certainly possible, we point out that in the context of histogram matching this would imply altering their semantic value (e.g. consider a mixture of two well-separated components representing different tissue types).

4.5.3 Warping

After matching one GMM to another, we also need a way to transform the data modelled by that GMM so it matches the target data. To this end, we borrow ideas from differential normalising flows (Section 2.2.3.2) and define the warping transformation, $\mathbf{f}$, as the trajectories of particles under the effect of a velocity field $\mathbf{u}$ over time t. Here, the temporal evolution of the density $q^{(t)}$ is specified directly from the density matching in Section 4.5.2, and we seek a velocity field $\mathbf{u}$ that satisfies *conservation of mass* (Eq. (2.33)). This section presents general results that apply to warping multivariate densities, though our intensity normalisation application involves only the univariate special case.

Let us first consider the case of warping a single mixture component. We begin by showing how to define mass-conserving velocity fields for reparametrised densities, such that samples continue to agree with them as parameters vary over time.

Lemma 4.2. *The velocity field* $(\mathbf{u}_t)_t$ *defined as*

$$\mathbf{u}^{(t)} := \dot{\boldsymbol{\psi}}_{\theta^{(t)}} \circ \boldsymbol{\psi}^{-1}_{\theta^{(t)}} = \langle \nabla_\theta \boldsymbol{\psi}_{\theta^{(t)}}, \dot{\theta}^{(t)} \rangle \circ \boldsymbol{\psi}^{-1}_{\theta^{(t)}} \tag{4.7}$$

conserves mass for a density $(q_{\theta^{(t)}})_t$ *built via reparametrisation* $(\boldsymbol{\psi}_{\theta^{(t)}})_t$ *of a fixed density* $\widetilde{q}$.

Proof. With $\mathbf{u}$ defined as above (notation simplified here for clarity), its divergence is expressed as

$$\nabla \cdot \mathbf{u} = \nabla \cdot (\dot{\boldsymbol{\psi}} \circ \boldsymbol{\psi}^{-1}) = \mathrm{tr}[(\nabla \boldsymbol{\psi})^{-1} \nabla \dot{\boldsymbol{\psi}}] \circ \boldsymbol{\psi}^{-1}.$$

Let $q = |\nabla \boldsymbol{\psi}^{-1}| \, \widetilde{q} \circ \boldsymbol{\psi}^{-1}$, such that $q \circ \boldsymbol{\psi} = |\nabla \boldsymbol{\psi}^{-1}| \, \widetilde{q}$. The total time derivative of q is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}(q \circ \boldsymbol{\psi}) = \frac{\partial q}{\partial t} \circ \boldsymbol{\psi} + (\nabla q \circ \boldsymbol{\psi}) \cdot \frac{\partial \boldsymbol{\psi}}{\partial t}.$$

Rearranging the terms and composing both sides with $\boldsymbol{\psi}^{-1}$ (which is valid because the reparametrisation $\boldsymbol{\psi}$ is assumed bijective), we can write

$$\begin{aligned}
\frac{\partial q}{\partial t} &= \left\{ \frac{\mathrm{d}}{\mathrm{d}t}\left(|\nabla \boldsymbol{\psi}|^{-1} \widetilde{q} \right) - (\nabla q \circ \boldsymbol{\psi}) \cdot \frac{\partial \boldsymbol{\psi}}{\partial t} \right\} \circ \boldsymbol{\psi}^{-1} \\
&= \left\{ - \underbrace{\left(|\nabla \boldsymbol{\psi}|^{-1} \widetilde{q} \right)}_{q \circ \boldsymbol{\psi}} \underbrace{\mathrm{tr}[(\nabla \boldsymbol{\psi})^{-1} \nabla \dot{\boldsymbol{\psi}}]}_{(\nabla \cdot \mathbf{u}) \circ \boldsymbol{\psi}} - (\nabla q \circ \boldsymbol{\psi}) \cdot \underbrace{\dot{\boldsymbol{\psi}}}_{\mathbf{u} \circ \boldsymbol{\psi}} \right\} \circ \boldsymbol{\psi}^{-1}
\end{aligned}$$

$$= -(q\nabla\cdot\mathbf{u} + \nabla q\cdot\mathbf{u}) = -\nabla\cdot(q\mathbf{u}),$$

where we have applied Jacobi's formula, $\partial_t|\nabla\boldsymbol{\psi}|^{-1} = -|\nabla\boldsymbol{\psi}|^{-1}\operatorname{tr}[(\nabla\boldsymbol{\psi})^{-1}\nabla\dot{\boldsymbol{\psi}}]$, and used the fact that $\partial_t\widetilde{q} = 0$. Hence $\mathbf{u} := \dot{\boldsymbol{\psi}}\circ\boldsymbol{\psi}^{-1}$ conserves mass for q reparametrised by $\boldsymbol{\psi}$. ■

Now, a random variable $x \sim \mathcal{N}(\mu_k, \lambda_k^{-1})$ can be expressed via a diffeomorphic reparametrisation of a standard Gaussian, with $x = \psi_k(\epsilon) = \mu_k + \epsilon/\sqrt{\lambda_k}$ and $\epsilon \sim \mathcal{N}(0,1)$. Assuming its mean and precision are changing with instantaneous rates $\dot{\mu}_k$ and $\dot{\lambda}_k$, respectively, we can introduce a velocity field $u_k = \dot{\psi}_k\circ\psi_k^{-1}$ for its samples so that they agree with this evolving density. The instantaneous velocity at 'time' t is thus given by

$$u_k^{(t)}(x) = \dot{\mu}_k^{(t)} - \frac{\dot{\lambda}_k^{(t)}}{2\lambda_k^{(t)}}\big(x - \mu_k^{(t)}\big). \tag{4.8}$$

In the case of a mixture with constant weights $\{\pi_k\}_k$, we can construct a smooth, *mass-conserving* global velocity field $\mathbf{u}$ as a point-wise convex combination of each component's velocity field, $\mathbf{u}_k$, weighted by the corresponding posterior assignment probabilities.

Proposition 4.3. *Let $(q_k)_k$ denote density functions and $(\mathbf{u}_k)_k$ velocity fields, and $(\pi_k)_k$ such that $\pi_k > 0$ and $\sum_k \pi_k = 1$. If the flow determined by each $\mathbf{u}_k$ conserves mass for the evolution of the respective q_k, then the flow determined by*

$$\mathbf{u}^{(t)}(x) := \sum_k \frac{\pi_k q_k^{(t)}(x)}{q^{(t)}(x)}\,\mathbf{u}_k^{(t)}(x), \tag{4.9}$$

conserves mass for the evolution of $q := \sum_k \pi_k q_k$.

Proof. Let us assume that, for some choice of parametric density family q_k, we have obtained a velocity field $\mathbf{u}_k$ that satisfies the continuity equation for the evolution of q_k, for each k:

$$\forall k, \quad \frac{\partial q_k}{\partial t} + \nabla\cdot(q_k\mathbf{u}_k) = 0. \tag{4.10}$$

Now, taking a convex combination of the above with weights $(\pi_k)_k$, we obtain

$$\begin{aligned} 0 = \sum_k \pi_k\left[\frac{\partial q_k}{\partial t} + \nabla\cdot(q_k\mathbf{u}_k)\right] &= \frac{\partial}{\partial t}\left(\sum_k \pi_k q_k\right) + \nabla\cdot\left(\sum_k \pi_k q_k\mathbf{u}_k\right) \\ &= \frac{\partial}{\partial t}\underbrace{\left(\sum_k \pi_k q_k\right)}_{q} + \nabla\cdot\Bigg[\underbrace{\left(\sum_k \pi_k q_k\right)}_{q}\underbrace{\left(\sum_k \frac{\pi_k q_k}{\sum_l \pi_l q_l}\mathbf{u}_k\right)}_{\sum_k \frac{\pi_k q_k}{q}\mathbf{u}_k =: \mathbf{u}}\Bigg] = \frac{\partial q}{\partial t} + \nabla\cdot(q\mathbf{u}). \end{aligned}$$ ■

Finally, the warping transformation $\mathbf{f}^{(t)}$ is given by the solution to the following ordinary differential equation (ODE):

$$\partial_t\mathbf{f}^{(t)}(x) = \mathbf{u}^{(t)}(\mathbf{f}^{(t)}(x)), \qquad \mathbf{f}^{(0)}(x) = x. \tag{4.11}$$

With $\mathbf{f}$ defined as above, we can prove that $q^{(t)}$ is indeed the density of samples from $q^{(0)}$ transformed through $\mathbf{f}^{(t)}$, i.e. $q^{(0)} = |\nabla \mathbf{f}^{(t)}|\, q^{(t)} \circ \mathbf{f}^{(t)}$ (Villani, 2009, p. 15). Crucially, the true solution to Eq. (4.11) is diffeomorphic by construction, and can be numerically approximated (and inverted) with arbitrary precision, along with its Jacobian log-determinant $\log|\nabla \mathbf{f}|$.

To this end, we propose to employ the classical fourth-order Runge–Kutta integration method (RK4), with step size η (0.01 in our experiments). At each step, following auxiliary values are computed:

$$\mathbf{k}_1 = \mathbf{u}^{(t)}(\mathbf{f}^{(t)}), \qquad k'_1 = \nabla \cdot \mathbf{u}^{(t)}(\mathbf{f}^{(t)}), \tag{4.12a}$$

$$\mathbf{k}_2 = \mathbf{u}^{(t+\eta/2)}(\mathbf{f}^{(t)} + \tfrac{\eta}{2}\mathbf{k}_1), \qquad k'_2 = \nabla \cdot \mathbf{u}^{(t+\eta/2)}(\mathbf{f}^{(t)} + \tfrac{\eta}{2}\mathbf{k}_1), \tag{4.12b}$$

$$\mathbf{k}_3 = \mathbf{u}^{(t+\eta/2)}(\mathbf{f}^{(t)} + \tfrac{\eta}{2}\mathbf{k}_2), \qquad k'_3 = \nabla \cdot \mathbf{u}^{(t+\eta/2)}(\mathbf{f}^{(t)} + \tfrac{\eta}{2}\mathbf{k}_2), \tag{4.12c}$$

$$\mathbf{k}_4 = \mathbf{u}^{(t+\eta)}(\mathbf{f}^{(t)} + \eta\mathbf{k}_3), \qquad k'_4 = \nabla \cdot \mathbf{u}^{(t+\eta)}(\mathbf{f}^{(t)} + \eta\mathbf{k}_3). \tag{4.12d}$$

Finally, we update the variables of interest:

$$\mathbf{f}^{(t+\eta)} \approx \mathbf{f}^{(t)} + \tfrac{\eta}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4), \tag{4.13a}$$

$$\log|\nabla \mathbf{f}^{(t+\eta)}| \approx \log|\nabla \mathbf{f}^{(t)}| + \tfrac{\eta}{6}(k'_1 + 2k'_2 + 2k'_3 + k'_4). \tag{4.13b}$$

Each update has $\mathcal{O}(\eta^5)$ local truncation error, producing a cumulative error of $\mathcal{O}(\eta^4)$, which is extremely accurate for moderate choices of step size η.

Now assume we obtain optimal parameter values $\{\mu_k^*\}_k$ and $\{\lambda_k^*\}_k$ after matching q to p. We can then warp the data using the above approach, for example linearly interpolating the intermediate parameter values, $\mu_k^{(t)} := t\mu_k^* + (1-t)\mu_k^{(0)}$ and $\lambda_k^{(t)} := t\lambda_k^* + (1-t)\lambda_k^{(0)}$, hence setting the rates in Eq. (4.8) to constant values, $\dot{\mu}_k = \mu_k^* - \mu_k^{(0)}$ and $\dot{\lambda}_k = \lambda_k^* - \lambda_k^{(0)}$, and integrating Eq. (4.11) for $t \in [0, 1]$.

4.5.4 Practical considerations

Since each medical image in a dataset can have millions of voxels, computing the posteriors and flows for every voxel individually can be too expensive for batch processing. To mitigate this issue, we can compute the end-to-end transformation on a mesh in the range of interest, which is then interpolated for the intensities in the entire volume. In the reported experiments, we have used a uniformly-spaced mesh of 200 points, which has proven accurate enough for normalisation purposes.

Note that the transformation could also be computed on the histogram of discrete intensity values and built into a look-up table. However, this would not scale well to two or more dimensions for multi-modal intensity normalisation, whereas a mesh would not need to be very fine nor require a

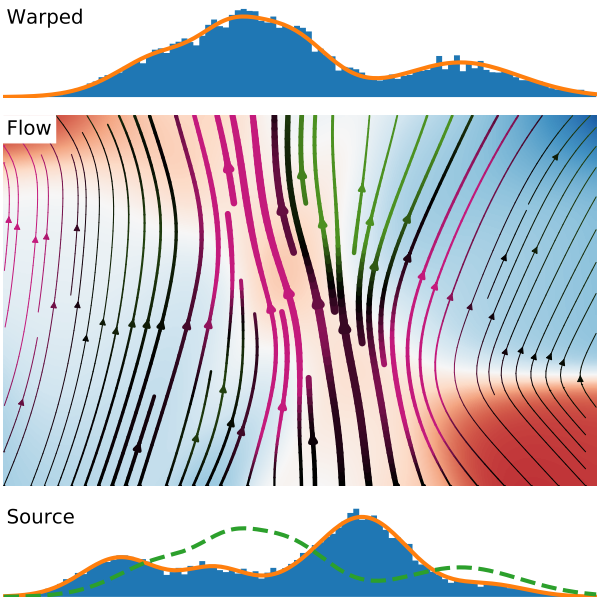


Figure 4.6 Example density flow between two histograms. Time flows upwards, from source histogram to warped histogram. The thickness of the streamlines indicates the local density, and their colour shows compression (pink, Jacobian > 1) and expansion (green, Jacobian < 1). The background colour represents the velocity field: red is negative (to the left), blue is positive (to the right).

regular grid layout.

4.6 Normalisation experiments

4.6.1 Dataset

Our experiments were run on 581 T1-weighted MRI scans from the IXI database, collected from three imaging centres with different scanners.[27] Each scan was bias field-corrected using SPM12[28] with default settings and rigidly registered to MNI space. SPM12 was further used to produce grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF) tissue probability maps. We obtained brain masks by adding the three probability maps and thresholding at 0.5. The statistics reported below were weighted by the voxel-wise tissue probabilities to account for partial-volume effects and segmentation ambiguities.

4.6.2 Setup

We firstly fitted the nonparametric mixture models to the full integer-value histograms of the raw images (inside the brain masks), as described in Section 4.5.1. We set the DP's concentration parameter to 2 and used data-driven Normal–Gamma priors for the components. As an ad-hoc post-processing step, we pruned the leftover mixture components with weights smaller than 10^{-3}. In the absence of one global reference distribution, we affinely aligned these DPGMMs and the corresponding data to

[27] http://brain-development.org/ixi-dataset/
[28] http://www.fil.ion.ucl.ac.uk/spm/software/spm12/

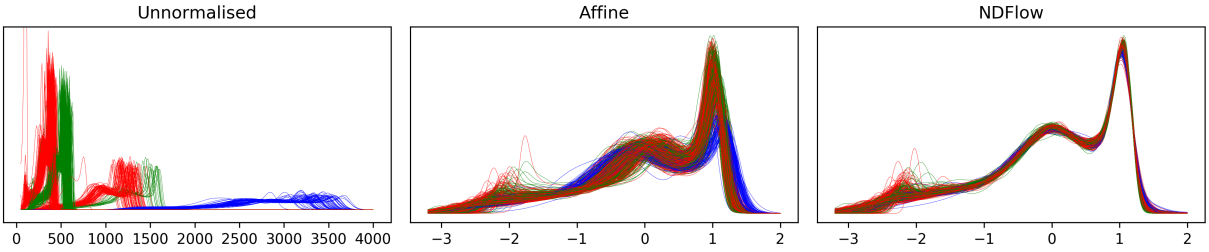
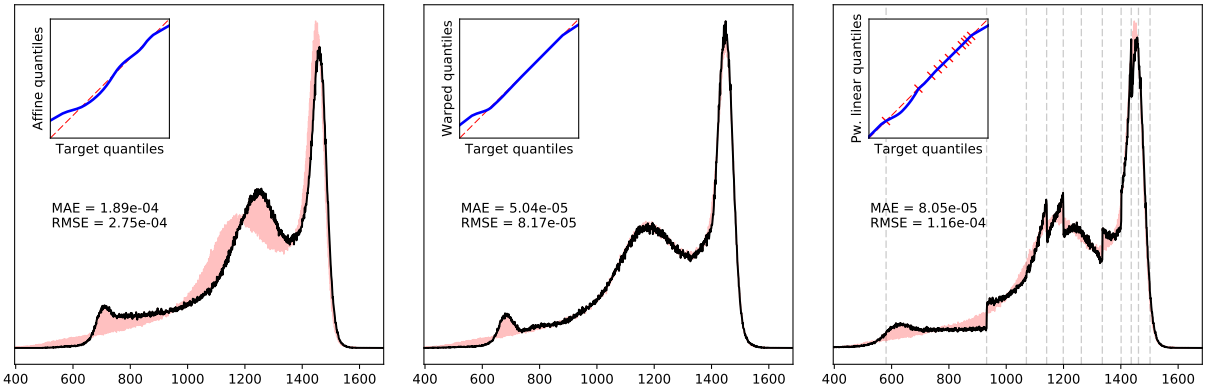


Figure 4.7 Intensity distributions, colour-coded by imaging centre



a Affine **b** NDFlow **c** Nyúl

Figure 4.8 Histograms and Q–Q plots of each of the methods against the target histogram. The shading shows the discrepancy between the transformed (black) and target histogram (light red). In the rightmost plot, the landmarks are indicated by vertical lines in the histogram and ticks in the Q–Q plot.

zero mean and unit variance (cf. Fig. 4.7, middle).

After this rough alignment, global and centre-wise average densities were computed. These were then considered as histograms to which we fitted global and centre-wise reference DPGMMs.

For normalisation, we consider two scenarios. The first is to normalise each centre's reference distribution to the global target, then to apply this same transformation to all subjects in that centre. In the other approach, each subject's image is individually normalised to the global target density. These scenarios reflect different practical applications where the centre-wise normalisation aims to preserve intra-centre variation, which might be desired. On the other hand, the individual normalisation aims to make all scans as similar as possible.

We compare our technique to Nyúl et al. (2000)'s prevalent quantile-based, piecewise linear histogram matching method, considered state-of-the-art for intensity normalisation and referred here as Nyúl. We acquired the default 11 landmarks (histogram deciles and upper/lower percentiles) from the affine-aligned data for all subjects, then normalised each subject to this set of average landmarks.

Table 4.2 Tissue statistics after normalisation. Mean ± std. dev., $N = 581$.

	Method	1st Quartile	Median	3rd Quartile
WM	Affine	0.900 ± 0.040	1.024 ± 0.045	1.126 ± 0.055
	NDFlow (Centre)	0.898 ± 0.040	1.020 ± **0.040**	1.121 ± **0.043**
	NDFlow (Individual)	0.890 ± **0.029**	1.014 ± **0.018**	1.120 ± **0.016**
	Nyúl	0.897 ± **0.029**	1.023 ± **0.015**	1.126 ± **0.008**
GM	Affine	−0.296 ± 0.142	0.025 ± 0.117	0.344 ± 0.080
	NDFlow (Centre)	−0.297 ± **0.139**	0.025 ± **0.114**	0.344 ± **0.076**
	NDFlow (Individual)	−0.312 ± **0.094**	0.027 ± **0.065**	0.351 ± **0.058**
	Nyúl	−0.309 ± **0.106**	0.027 ± **0.070**	0.350 ± **0.064**
CSF	Affine	−2.036 ± 0.145	−1.486 ± 0.140	−1.024 ± 0.156
	NDFlow (Centre)	−2.035 ± **0.143**	−1.480 ± 0.142	−1.018 ± 0.160
	NDFlow (Individual)	−2.031 ± **0.136**	−1.484 ± 0.170	−1.028 ± 0.191
	Nyúl	−2.025 ± **0.111**	−1.474 ± 0.178	−1.029 ± 0.207

Bold: $p < .01$, one-tailed Brown–Forsythe test for lower variance than 'Affine'

4.6.3 Histogram fitness

Figure 4.8 illustrates the results of normalisation between the pair of images in Fig. 4.5, which have a notable dissimilarity in the CSF region of the histograms. We observe that both our NDFlow- and Nyúl-transformed histograms present substantially lower mean absolute and root mean squared errors (MAE and RMSE) than the affine-aligned one, and our method performed best by a small margin. This is confirmed in a number of trials with other images.

A noteworthy artefact of Nyúl are abrupt jumps produced at the landmark values (e.g. Fig. 4.8c), which appear because intervals are uniformly compressed or dilated by different factors, and may be detrimental to downstream histogram-based tasks (e.g. mutual information registration). NDFlow causes no such discontinuities due to the smoothness of the mass-conserving flows.

4.6.4 Tissue statistics

In Table 4.2 we report the WM, GM and CSF intensity statistics for different normalisations. Firstly, we see that the centre-wise normalisation had a small but significant effect on the overall distribution statistics. More importantly, the variances of the statistics after individual NDFlow and Nyúl transformations were typically similar, and both were almost always substantially smaller than the variance after only affine alignment, with the exception of CSF.

It is known that the amount of intra-cranial fluid can vary substantially due to factors such as age and some neurodegenerative conditions, and this reflects on the distributions of intensities in brain MRI scans, which is evident in Fig. 4.7. As a result, normalising all subjects to a 'mean' distribution

Table 4.3 Centre classification accuracy. The central columns correspond to the three London imaging centres where IXI data was collected: Guy's Hospital, Hammersmith Hospital (HH) and Institute of Psychiatry (IOP). The 'Overall' column shows the class-balanced average accuracy.

	Guy's	HH	IOP	Overall
Unnormalised	0.9906	0.9892	1.0000	0.9913
Affine	0.9687	0.9135	0.8904	0.9411
NDFlow (Centre)	0.9687	0.9081	0.7945	0.9272
NDFlow (Individual)	0.9749	0.9135	0.8219	0.9359
Nyúl	0.9655	0.9189	0.7945	0.9289

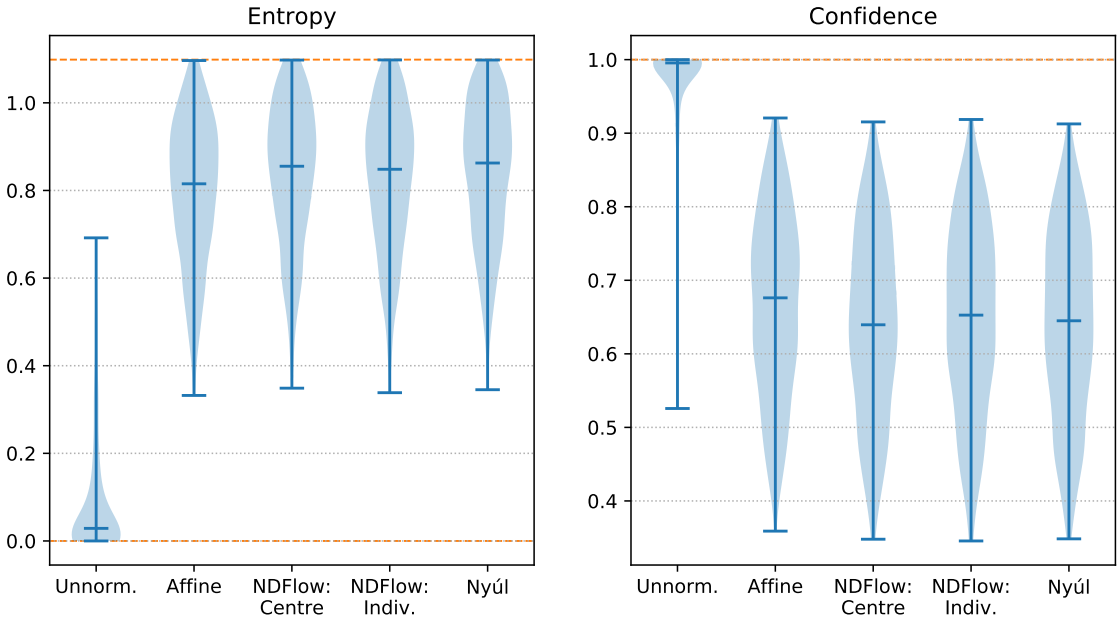


Figure 4.9 Centre prediction statistics. Each violin plot indicates minimum, maximum and median. Dashed horizontal lines mark maximal possible values ($\log 3 \approx 1.10$ for entropy). Confidence here is the predicted probability of the chosen class.

fails to identify a consistent reference range for CSF intensities.

A fundamental limitation of any histogram matching scheme is that it is unclear how to proceed when the distributions are *genuinely* different. Intensity distributions can be strongly affected by anatomical differences; for example, we can observe large variations in the amounts of fluid and fat in brain or whole-body scans, which may heavily skew the overall distributions (moderate example in Fig. 4.8). The underlying assumption of these methods (including ours) is that the distributions are similar enough up to an affine rescaling and a mild nonlinear deformation of the values, thus handling histograms of truly different shapes remains an open challenge. For images with different fields of view, it may be beneficial to perform image registration before applying intensity normalisation.

4.6.5 Centre classification

To evaluate the effectiveness of intensity normalisation for data harmonisation, we conducted a centre discrimination experiment with random forest classifiers trained on the full images. We report the pooled test results from two-fold cross validation.

Relative to affine normalisation, centre-wise and individual NDFLOW and NYÚL showed a slight drop in overall classification accuracy (94.1% vs. 92.7%, 93.6%, 92.9%, resp.). On the other hand, the uncertainty, as measured by the entropy of the predictions, was significantly higher (paired t-test, all $p < .01$). Nonlinear intensity normalisation therefore seems to successfully remove some of the biasing factors which are discriminative of the origin of the images, which relates to the findings in Glocker et al. (2019).

4.7 Conclusion

To the best of our knowledge, the initial part of this chapter is the first application of nonparametric mixture models to multimodal medical image segmentation. Under our experimental conditions, we have observed that leveraging multiple imaging modalities drives the DPMM towards better segmentations than those obtained with a single modality, which was confirmed in terms of both labelling consistency and spatial accuracy.

A further contribution of this work is that this is the first known attempt at applying DPMMs to lesion segmentation in medical images. We have shown that, at least in the 2D case, this nonparametric approach finds a good correspondence to the manual segmentations for large enough lesions. Additionally, we have verified that DPGMMs are able to provide convincing representations of multidimensional probability distributions.

In the second part, we have introduced a novel method for MRI intensity normalisation, called *nonparametric density flows* (NDFLOW). It is based on fitting and matching DPGMM densities, by minimising their L^2 divergence, and on mass-conserving flows, which ensure that the empirical intensity distribution agrees with the matched density model.

We demonstrated that our normalisation approach makes tissue intensity statistics significantly more consistent across subjects than a simple affine alignment, and compares favourably to the state-of-the-art method of Nyúl et al. (2000). We have additionally verified that NDFLOW is able to accurately match histograms without introducing spurious artefacts produced by the competing method. Finally, we argued that both normalisation techniques can reduce some discriminative scanner biases, in a step toward effective data harmonisation for mitigating acquisition shift (Section 3.4.1).

This chapter focused only on image histograms, and the second part additionally considered their

relationship with acquisition centre labels. To demonstrate more complex modelling strategies we take a detour through a different application domain involving imaging: face recognition. The methodology presented in Chapter 5 illustrates how some assumptions about a problem domain can be naturally translated into the formulation of a probabilistic model. In particular, we leverage the flexibility of nonparametric models (as employed in the present chapter) to capture the unboundedness of possible identities, names, and situational contexts. The next chapter also exemplifies how different types of observed data can be plugged into a larger model by means of appropriate likelihood models: e.g. images via lower-dimensional embeddings, sparse and noisy name labels with some simple language model, and environmental cues following spatial or temporal distributions.

Chapter 5

A probabilistic model for contextual face recognition

This chapter is based on the following publications:

- **Daniel C. Castro** and Sebastian Nowozin (2018b). 'From face recognition to models of identity: a Bayesian approach to learning about unknown identities from unsupervised data'. In: *Computer Vision – ECCV 2018* (Munich, Germany). Vol. 11206. LNCS. Springer, Cham, pp. 764–780. DOI: 10.1007/978-3-030-01216-8_46. arXiv: 1807.07872;
- **Daniel C. Castro** and Sebastian Nowozin (2018a). 'Contextual face recognition with a nested-hierarchical nonparametric identity model'. In: *NeurIPS 2018 Workshop on All of Bayesian Nonparametrics* (Montréal, Canada). arXiv: 1811.07753,

and patent:

- Sebastian Nowozin, Tom Ellis, Cecily Morrison and **Daniel C. Castro** (2019). 'Image Processing for Person Recognition'. U.S. patent 10621416. Microsoft Technology Licensing, LLC.

This work was conducted during a research internship at Microsoft Research, Cambridge, UK.

The author emphasises he is strictly opposed to the use of automatic facial recognition for mass surveillance. The research presented here is intended for an assistive technology that empowers individuals who are blind or with low vision to better perceive their surroundings, and the development of the overarching project gave great emphasis to ethics, consent, and privacy. More details can be found at https://www.microsoft.com/en-us/research/project/project-tokyo/ and https://news.microsoft.com/innovation-stories/project-tokyo/ (accessed 23/09/20).

Current face recognition systems robustly recognise identities across a wide variety of imaging conditions. In these systems, recognition is performed via classification of face images into known identities obtained from supervised name annotations. There are two problems with this current paradigm: (1)

current systems are unable to benefit from unlabelled data, which may be available in large quantities; and (2) current systems equate successful recognition with labelling a given input image. Humans, on the other hand, regularly perform identification of individuals completely unsupervised, recognising the identity of someone they have seen before even without being able to name that individual. How can we surpass the current classification paradigm to achieve a more human understanding of identities?

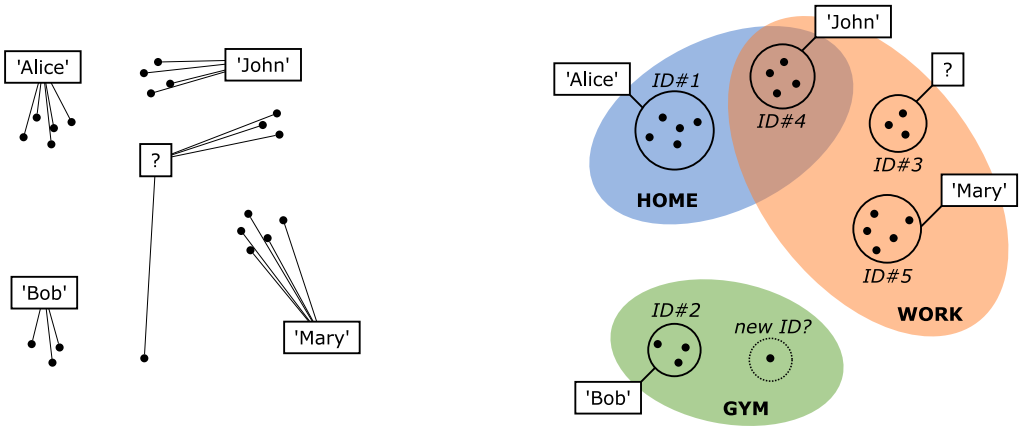
We propose an integrated Bayesian model that coherently reasons about the observed images, identities, partial knowledge about names, and the situational context of each observation. While our model achieves good recognition performance against known identities, it can also discover new identities from unlabelled data and learns to associate identities with different contexts depending on which identities tend to be observed together. In addition, the proposed semi-supervised component is able to handle not only acquaintances, whose names are known, but also unlabelled familiar faces and complete strangers in a unified framework. At the end of the chapter, we propose extensions of the contextual component of this model, enabling unsupervised discovery of an unbounded number of contexts for improved face recognition.

5.1 Introduction

For the following discussion, we decompose the usual face identification task into two sub-problems: *recognition* and *tagging*. Here we understand recognition as the unsupervised task of matching an observed face to a cluster of previously seen faces with similar appearance (disregarding variations in pose, illumination etc.), which we refer to as an *identity*. Humans routinely operate at this level of abstraction to recognise familiar faces: even when people's names are not known, we can still tell them apart. Tagging, on the other hand, refers to putting names to faces, i.e. associating string literals to known identities.

Humans tend to create an inductive mental model of facial appearance for each person we meet, which we then query at new encounters to be able to recognise them. This is opposed to a transductive approach, attempting to match faces to specific instances from a memorised gallery of past face observations—which is how identification systems are often implemented (Jafri and Arabnia, 2009).

An alternative way to represent faces, aligned with our inductive recognition, is via *generative* face models, which explicitly separate latent identity content, tied across all pictures of a same individual, from nuisance factors such as pose, expression, and lighting (Ioffe, 2006; P. Li et al., 2012; Prince and Elder, 2007). While mostly limited to linear projections from pixel space (or mixtures thereof), the probabilistic framework applied in these works allowed tackling a variety of face recognition tasks,



a Standard face recognition **b** Context-aware model of identities

Figure 5.1 Conventional and proposed face recognition paradigms. Points represent face observations and boxes are name labels. Rather than attempting to label each face image in isolation, the proposed perspective groups observations of the same individual into an identity, with an associated name label that may be unknown. In addition, it acknowledges that identities are encountered in different contexts, which can help narrow down the possibilities when recognising a new observation.

such as closed- and open-set identification, verification and clustering.

A further important aspect of social interactions is that, as an individual continues to observe faces every day, they encounter some people much more often than others, and the total number of distinct identities ever met tends to increase virtually without bounds. Additionally, we argue that human face recognition does not happen in an isolated environment, but situational contexts (e.g. 'home', 'work', 'gym') constitute strong cues for the groups of people a person expects to meet (Fig. 5.1b).

With regards to tagging, in daily life we very rarely obtain named face observations: acquaintances normally introduce themselves only once, and not repeatedly whenever they are in our field of view. In other words, humans are naturally capable of semi-supervised learning, generalising sparse name annotations to all observations of the corresponding individuals, while additionally reconciling naming conflicts due to noise and uncertainty.

In contrast, standard computational face identification is *fully supervised* (see Fig. 5.1a), relying on vast labelled databases of high-quality images (Learned-Miller and G. B. Huang, 2016). Although many supervised methods achieve astonishing accuracy on challenging benchmarks (e.g. Schroff et al., 2015; Taigman et al., 2014) and are successfully employed in practical biometric applications, this setting has arguably limited analogy to human social experience.

Expanding on the generative perspective, we introduce a unified Bayesian model which reflects all the above considerations on identity distributions, context-awareness and labelling (Fig. 5.1b). Our nonparametric identity model effectively represents an unbounded population of identities, while taking contextual co-occurrence relations into account and exploiting modern deep face representa-

tions to overcome limitations of previous linear generative models. Our main contributions in this work are twofold:

1. We propose an unsupervised face recognition model which can explicitly reason about people it has never seen; and

2. We attach to it a novel robust label model enabling it to predict names by learning from both named and unnamed faces.

5.2 Related work

Other face recognition methods (even those formulated in a Bayesian framework), often limit themselves to point estimates of parameters and predictions, occasionally including ad-hoc confidence metrics (Choi et al., 2010; Le et al., 2017; Tapaswi et al., 2012; Lei Zhang et al., 2003; M. Zhao et al., 2006). A distinct advantage of our approach is that it is probabilistic end-to-end, and thus naturally provides predictions with principled, quantifiable uncertainties. Moreover, we employ modern Bayesian modelling tools—namely hierarchical nonparametrics—which enable dynamically adapting model complexity while faithfully reflecting the real-world assumptions laid out above.

Secondly, although automatic face tagging is a very common task, each problem setting can impose wildly different assumptions and constraints. Typical application domains involve the annotation of personal photo galleries (Anguelov et al., 2007; Gallagher and T. Chen, 2009; Lei Zhang et al., 2003; M. Zhao et al., 2006), multimedia (e.g. TV; Le et al., 2017; Tapaswi et al., 2012) or security/surveillance (Jafri and Arabnia, 2009). Our work focuses on egocentric human-like face recognition, a setting which seems largely unexplored, as most of the work using first-person footage appears to revolve around other tasks like object and activity recognition, face detection, and tracking (Betancourt et al., 2015). As we explained previously, the dynamic, *online* nature of first-person social experience brings a number of specific modelling challenges for face recognition.

Finally, while there is substantial prior work on using contexts to assist face recognition, we emphasise that much (perhaps most) of it is effectively complementary to our unified framework. Notions of *global* context such as timestamp, geolocation and image background (Anguelov et al., 2007; Choi et al., 2010; Torralba et al., 2003; M. Zhao et al., 2006) can readily be used to inform our current context model (Section 5.3.1). In addition, we can naturally augment the proposed face model (Section 5.3.3) to leverage further *individual* context features, e.g. clothing and speech (Anguelov et al., 2007; Le et al., 2017; Tapaswi et al., 2012; M. Zhao et al., 2006). Integration of these additional factors opens exciting avenues for future research.

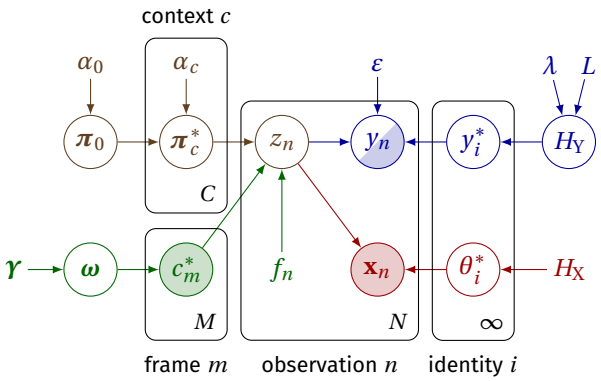


Figure 5.2 Overview of the proposed generative model, encompassing the *context model*, *identity model*, *face model*, and *label model*. Unfilled nodes represent latent variables, shaded nodes are observed, the half-shaded node is observed only for a subset of the indices and uncircled nodes denote fixed hyperparameters. $\boldsymbol{\pi}_0$ and $(\boldsymbol{\pi}_c^*)_{c=1}^C$ are the global and context-wise identity probabilities; $\boldsymbol{\omega}$ denotes the context probabilities; $(c_m^*)_{m=1}^M$ are the frame-wise context labels, indexed by the frame numbers $(f_n)_{n=1}^N$; $(z_n)_{n=1}^N$ are the latent identity indicators; $(\mathbf{x}_n)_{n=1}^N$ are the face observations and $(y_n)_{n=1}^N$ are the respective name annotations; $(\theta_i^*)_{i=1}^\infty$ are the parameters of the face model; and $(y_i^*)_{i=1}^\infty$ are the identities' name labels. See text for descriptions of the remaining symbols.

5.3 A model of identities

In this section, we describe in isolation each of the building blocks of the proposed approach to facial identity recognition: the context model, the identity model and the face model. We assume data is collected in the form of camera *frames* (either photographs or video stills), numbered 1 to M, and faces are cropped with some face detection system and grouped by frame number indicators, $f_n \in \{1,\dots,M\}$. The diagram in Fig. 5.2 illustrates the full proposed graphical model, including the label model detailed in Section 5.4.

5.3.1 Context model

In our identity recognition scenario, we imagine the user moving between contexts throughout the day (e.g. home–work–gym…). Since humans naturally use situational context as a strong prior on the groups of people we expect to encounter in each situation, we incorporate context-awareness in our model of identities to mimic human-like face recognition.

The context model we propose involves a discrete variable $c_n \in \{1,\dots,C\}$ for each observation, where C is some fixed number of distinct contexts, representing categories of situations in which a subject may find herself: e.g. home, work, gym.[29] Crucially, we assume that all face detections in frame m, $\mathcal{F}_m = \{n : f_n = m\}$, share the same context, c_m^* (i.e. $\forall n, c_n = c_{f_n}^*$).

We define the identity indicators to be independent given the context of the corresponding frames (see Section 5.3.2, below). However, since the contexts are tied by frame, marginalising over the

[29] We explore an extension to unbounded contexts in Section 5.6.1.

contexts captures identity co-occurrence relations. In turn, these allow the model to make more confident predictions about people who tend to be seen together in the same environment.

This formalisation of contexts as discrete semantic labels is closely related to the place recognition model in Torralba et al. (2003), used there to disambiguate predictions for object detection. It has also been demonstrated that explicit incorporation of a context variable can greatly improve clustering with mixture models (Perdikis et al., 2015).

Finally, we assume the context indicators c^*_m are independently distributed according to probabilities $\boldsymbol{\omega}$, which themselves follow a Dirichlet prior:

$$\boldsymbol{\omega} \sim \mathrm{Dir}(\boldsymbol{\gamma}) \tag{5.1}$$

$$c^*_m \,|\, \boldsymbol{\omega} \sim \mathrm{Cat}(\boldsymbol{\omega}), \qquad m = 1, \ldots, M, \tag{5.2}$$

where M is the total number of frames. In our simulation experiments, we use a symmetric Dirichlet prior, setting $\boldsymbol{\gamma} = (\gamma_0/C, \ldots, \gamma_0/C)$.

5.3.2 Identity model

In the daily-life scenario described in Section 5.1, an increasing number of unique identities will tend to appear as more faces are observed. This number is expected to grow much more slowly than the number of observations, and can be considered unbounded in practice (we do not expect a user to run out of new people to meet). Moreover, we can expect some people to be encountered much more often than others. Since a Dirichlet process (DP; Ferguson, 1973) displays properties that mirror all of the above phenomena (Teh, 2010), it is a sound choice for modelling the distribution of identities.

Furthermore, the assumption that all people can potentially be encountered in any context, but with different probabilities, is perfectly captured by a hierarchical Dirichlet process (HDP; Teh et al., 2006). Making use of the context model, we define one DP *per context* c, each with concentration parameter α_c and sharing the same *global* DP as a base measure. This hierarchical construction thus produces context-specific distributions over a common set of identities.

We consider that each of the N face detections is associated to a latent identity indicator variable, z_n. Letting $\boldsymbol{\pi}_0$ denote the global identity distribution and $(\boldsymbol{\pi}^*_c)_{c=1}^C$ the context-specific identity distributions, we can write the generative process as

$$\boldsymbol{\pi}_0 \sim \mathrm{GEM}(\alpha_0) \tag{5.3}$$

$$\boldsymbol{\pi}^*_c \,|\, \boldsymbol{\pi}_0 \sim \mathrm{HGEM}(\alpha_c, \boldsymbol{\pi}_0), \qquad c = 1, \ldots, C \tag{5.4}$$

$$z_n \,|\, f_n = m, \mathbf{c}^*, (\boldsymbol{\pi}^*_c)_c \sim \mathrm{Cat}(\boldsymbol{\pi}^*_{c^*_m}), \qquad n = 1, \ldots, N, \tag{5.5}$$

where GEM(α_0) is the DP stick-breaking distribution, $\pi_{0i} = \beta_{0i}\prod_{j=1}^{i-1}(1-\beta_{0j})$, with $\beta_{0i} \sim \text{Beta}(1,\alpha_0)$ and $i = 1,\ldots,\infty$ (Pitman, 2006; Sethuraman, 1994). We additionally define a *hierarchical GEM distribution*, HGEM($\alpha, \boldsymbol{\pi}_0$), such that $\pi^*_{ci} = \beta_{ci}\prod_{j=1}^{i-1}(1-\beta_{cj})$, with $\beta_{ci} \sim \text{Beta}\left(\alpha_c\pi_{0i}, \alpha_c(1-\sum_{j=1}^{i}\pi_{0j})\right)$ (Teh et al., 2006, Eq. (21)).

Although the full generative model involves infinite-dimensional objects, DP-based models present simple finite-dimensional marginals. In particular, the posterior predictive probability of encountering a known identity i is

$$p(z_{N+1} = i \mid c_{N+1} = c, \mathbf{z}, \mathbf{c}^*, \boldsymbol{\pi}_0) = \frac{\alpha_c\pi_{0i} + N_{ci}}{\alpha_c + N_{c\cdot}}, \tag{5.6}$$

where N_{ci} is the number of observations assigned to context c and identity i and $N_{c\cdot}$ is the total number of observations in context c.

Finally, such a nonparametric model is well suited for an open-set identification task, as it can elegantly estimate the prior probability of encountering an unknown identity:

$$p(z_{N+1} = I+1 \mid c_{N+1} = c, \mathbf{z}, \mathbf{c}^*, \boldsymbol{\pi}_0) = \frac{\alpha_c\pi'_0}{\alpha_c + N_{c\cdot}}, \tag{5.7}$$

where I is the current number of distinct known identities and $\pi'_0 = \sum_{i=I+1}^{\infty}\pi_{0i}$ denotes the global probability of sampling a new identity.

5.3.3 Face model

We assume that the observed features of the n^{th} face, $\mathbf{x}_n$, arise from a parametric family of distributions, F_X. The parameters of this distribution, θ^*_i, drawn from a prior, H_X, are unique for each identity and are shared across all face feature observations of the same person:

$$\theta^*_i \sim H_X, \qquad i = 1,\ldots,\infty \tag{5.8}$$

$$\mathbf{x}_n \mid z_n, \boldsymbol{\theta}^* \sim F_X(\theta^*_{z_n}), \qquad n = 1,\ldots,N. \tag{5.9}$$

As a consequence, the marginal distribution of faces is given by an *infinite mixture model* (Antoniak, 1974): $p(\mathbf{x}_n \mid c_n = c, \boldsymbol{\theta}^*, \boldsymbol{\pi}^*_c) = \sum_{i=1}^{\infty}\pi^*_{ci}F_X(\mathbf{x}_n \mid \theta^*_i)$.

In face recognition applications, it is typically more convenient and meaningful to extract a compact representation of face features than to work directly in a high-dimensional pixel space. For the experiments reported in this chapter, we used the 128-dimensional embeddings produced by OpenFace, a publicly available, state-of-the-art neural network for face recognition (Amos et al., 2016), implementing FaceNet's architecture and methodology (Schroff et al., 2015). In practice, this could

easily be swapped for other face embeddings (e.g. DeepFace; Taigman et al., 2014) without affecting the remainder of the model. We chose isotropic Gaussian mixture components for the face features (F_X), with an empirical Gaussian–inverse gamma prior for their means and variances (H_X).

5.4 Semi-supervised label model

We expect to work with only a small number of labelled observations manually provided by the user. Since the final goal is to identify any observed face, our probabilistic model needs to incorporate a semi-supervised aspect, generalising the sparse given labels to unlabelled instances. Throughout this section, the terms 'identity' and 'cluster' will be used interchangeably.

One of the cornerstones of semi-supervised learning (SSL) is the premise that clustered items tend to belong to the same class (Chapelle et al., 2006, Sec. 1.2.2). Building on this *cluster assumption*, mixture models, such as ours, have been successfully applied to SSL tasks (Bouveyron and Girard, 2009). We illustrate in Fig. 5.3 our proposed label model detailed below, comparing it qualitatively to nearest-neighbour classification on a toy example.

With the motivation above, we attach a label variable (a *name*) to each cluster (identity), here denoted y^*_i. This notation suggests that there is a single true label $\tilde{y}_n = y^*_{z_n}$ for each observation n, analogously to the observation parameters: $\theta_n = \theta^*_{z_n}$. Finally, the observed labels, y_n, are potentially corrupted through some noise process, F_Y. Let $\mathcal{L}$ denote the set of indices of the labelled data. The complete generative process is presented below:

$$H_Y \sim \mathrm{DP}(\lambda, L) \tag{5.10}$$

$$y^*_i \mid H_Y \sim H_Y, \qquad i = 1, \ldots, \infty \tag{5.11}$$

$$y_n \mid z_n, \mathbf{y}^* \sim F_Y(y^*_{z_n}), \qquad n \in \mathcal{L}\,. \tag{5.12}$$

As mentioned previously, a related model for mixture model-based SSL with noisy labels was proposed in Bouveyron and Girard (2009). Instead of considering an explicit noise model for the class labels, the authors of that work model directly the conditional label distribution for each cluster. Our setting here is more general: we assume not only an unbounded number of clusters, but also of possible labels.

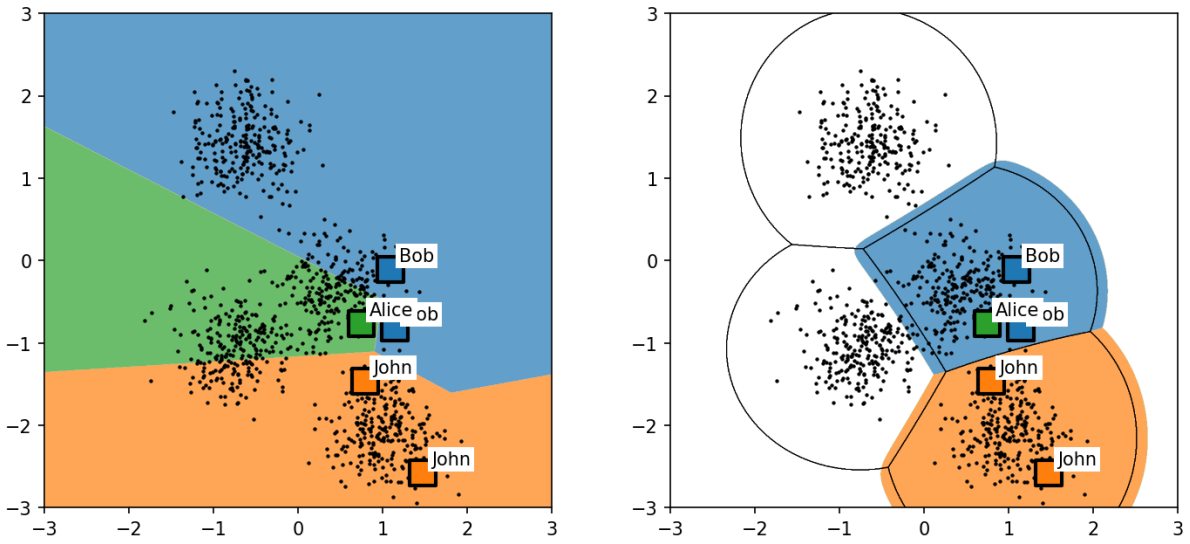


Figure 5.3 Hard label predictions of the proposed semi-supervised label model (right) and nearest-neighbour classification (left). Points represent unlabelled face observations, squares are labelled and the black contours on the right show identity boundaries. The proposed label model produces more natural boundaries, assigning the 'unknown' label (white) to unlabelled clusters and regions distant from any observed cluster, while also accommodating label noise ('Bob' → 'Alice') without the spurious boundaries introduced by NN.

5.4.1 Label prior

We assume that the number of distinct labels will tend to increase without bounds as more data is observed. Therefore, we adopt a further nonparametric prior on the cluster-wide labels:

$$H_{\mathrm{Y}} \sim \mathrm{DP}(\lambda, L)\,, \tag{5.13}$$

where L is some base probability distribution over the countable but unbounded label space (e.g. strings).[30] We briefly discuss the choice of L further below.

All concrete knowledge we have about the random label prior H_{Y} comes from the set of observed labels, $\mathbf{y}_{\mathcal{L}}$. Crucially, if we marginalise out H_{Y}, the predictive label distribution is simply (Teh, 2010)

$$y^*_{I+1} \,|\, \mathbf{y}^* \sim \frac{1}{\lambda + I}\Big(\lambda L + \sum_{\ell \in \mathcal{Y}} J_\ell \delta_\ell\Big), \tag{5.14}$$

which we will denote $\widehat{H_{\mathrm{Y}}}(y^*_{I+1} \,|\, \mathbf{y}^*)$. Here, $\mathcal{Y}$ is the set of distinct known labels among $\mathbf{y}_{\mathcal{L}}$ and $J_\ell = |\{i : y^*_i = \ell\}|$, the number of components with label ℓ (note that $\sum_\ell J_\ell = I$).

In addition to allowing multiple clusters to have repeated labels, this formulation allows us to reason about *unseen* labels. For instance, some of the learned clusters may have no labelled training points assigned to them, and the true (unobserved) labels of those clusters may never have been

[30] One could instead consider a Pitman–Yor process if power-law behaviour seems more appropriate than the DP's exponential tails (Pitman and Yor, 1997).

encountered among the training labels. Another situation in which unseen labels come into play is with points away from any clusters, for which the identity model would allocate a new cluster with high probability. In both cases, this model gives us a principled estimate of the probability of assigning a special 'unknown' label.

The base measure L may be defined over a rudimentary language model. For this work, we adopted a geometric/negative binomial model for the string length $|\ell|$, with characters drawn uniformly from an alphabet of size K:

$$L_{\phi,K}(\ell) = \mathrm{Geom}\left(|\ell|; \tfrac{1}{\phi}\right) \mathrm{Unif}\left(\ell; K^{|\ell|}\right) = \frac{1}{\phi - 1}\left(\frac{\phi - 1}{\phi K}\right)^{|\ell|}, \tag{5.15}$$

where ϕ is the expected string length.

5.4.2 Label likelihood

In the simplest case, we could consider $F_{\mathrm{Y}}(\cdot) = \delta_{\cdot}$, i.e. noiseless labels. Although straightforward to interpret and implement, this could make inference highly unstable whenever there would be conflicting labels for an identity. Moreover, in our application, the labels would be provided provided by a human user who may not have perfect knowledge of the target person's true name or its spelling, for example.

Therefore, we incorporate a label noise model, which can gracefully handle conflicts and mislabelling. We assume observed labels are noisy completely at random (NCAR; Frénay and Verleysen, 2014, Sec. II-C), with a fixed error rate ε:[31]

$$\widehat{F_{\mathrm{Y}}}(\ell \mid y_i^*; \mathbf{y}^*) = \begin{cases} 1 - \varepsilon, & \ell = y_i^* \\ \varepsilon \dfrac{\widehat{H_{\mathrm{Y}}}(\ell \mid \mathbf{y}^*)}{1 - \widehat{H_{\mathrm{Y}}}(y_i^* \mid \mathbf{y}^*)}, & \ell \neq y_i^* \end{cases}. \tag{5.16}$$

Intuitively, an observed label, y_n, agrees with its identity's assigned label, $y^*_{z_n}$, with probability $1 - \varepsilon$. Otherwise, it is assumed to come from a modified label distribution, in which we restrict and renormalise $\widehat{H_{\mathrm{Y}}}$ to exclude $y^*_{z_n}$. Here we use $\widehat{H_{\mathrm{Y}}}$ in the error distribution instead of L to reflect that a user is likely to mistake a person's name for another known name, rather than for a completely random string.

[31] The 'true' label likelihood $F_{\mathrm{Y}}(\ell \mid y_i^*)$ is random due to its dependence on the unobserved prior H_{Y}. We thus define $\widehat{F_{\mathrm{Y}}}$ as its posterior expectation given the known identity labels $\mathbf{y}^*$. See Appendix A.1 for details.

5.4.3 Label prediction

For label prediction, we are only concerned with the true, noiseless labels, $\tilde{y}_n$. The predictive distribution for a single new sample is given by

$$p(\tilde{y}_{N+1} = \ell \,|\, \mathbf{x}_{N+1}, \mathbf{z}, \mathbf{c}^*, \mathbf{y}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi}_0) = \sum_{i \leq I : y_i^* = \ell} p(z_{N+1} = i \,|\, \mathbf{x}_{N+1}, \mathbf{z}, \mathbf{c}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi}_0) + \widehat{H_Y}(y_{I+1}^* = \ell \,|\, \mathbf{y}^*)\; p(z_{N+1} = I+1 \,|\, \mathbf{x}_{N+1}, \mathbf{z}, \mathbf{c}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi}_0)\,. \tag{5.17}$$

The sum in the first term is the probability of the sample being assigned to any of the existing identities that have label ℓ, while the last term is the probability of instantiating a new identity with that label.

5.5 Evaluation

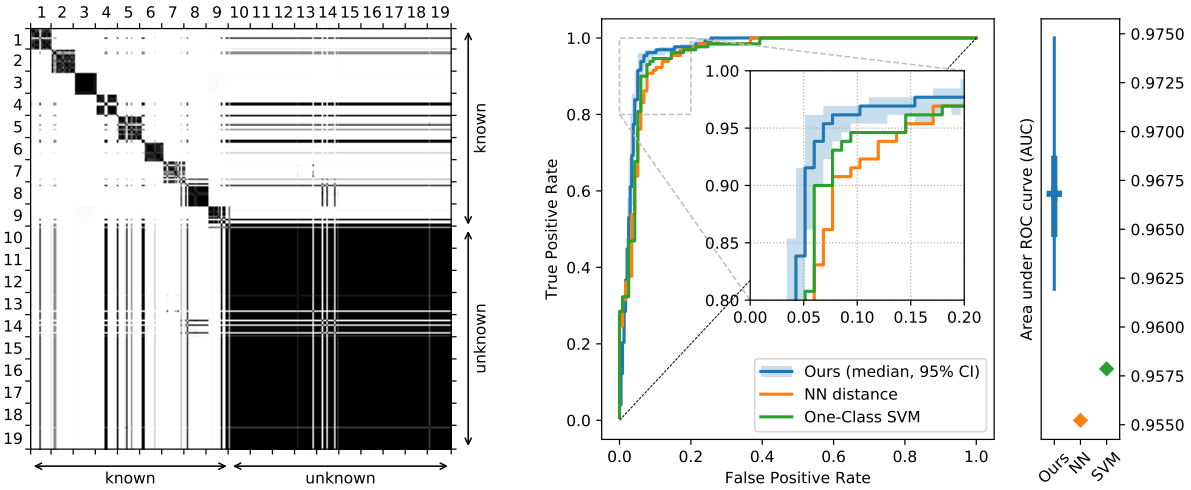
One of the main strengths of the proposed model is that it creates a single rich representation of the known world, which can then be queried from various angles to obtain distinct insights. In this spirit, we designed three experimental setups to assess different properties of the model: detecting whether a person has been seen before (outlier detection), recognising faces as different identities in a sequence of frames (clustering, unsupervised) and correctly naming observed faces by generalising sparse user annotations (semi-supervised learning).

In all experiments, we used celebrity photographs from the Labelled Faces in the Wild (LFW) database (G. B. Huang et al., 2007). We have implemented inference via Gibbs Markov chain Monte Carlo (MCMC) sampling, whose conditional distributions can be found in Appendix A.2, and we run multiple chains with randomised initial conditions to better estimate the variability in the posterior distribution. For all metrics evaluated on our model, we report the estimated 95% highest posterior density (HPD) credible intervals over pooled samples from 8 independent Gibbs chains, unless stated otherwise.

5.5.1 Experiment 1: Unknown person detection

In our first set of experiments, we study the model's ability to determine whether or not a person has been seen before. This key feature of the proposed model is evaluated based on the probability of an observed face not corresponding to any of the known identities, as given by Eq. (5.7). In order to evaluate purely the detection of unrecognised faces, we constrained the model to a single context ($C = 1$) and set aside the label model ($\mathcal{L} = \emptyset$).

This task is closely related to outlier/anomaly detection. In particular, our proposed approach mirrors one of its common formulations, involving a mixture of a 'normal' distribution, typically fitted



a Association matrix, counting agreements in the MAP identity predictions (including *unknown*). Ticks delimit ground-truth identities.

b ROC analysis for unknown person detection compared with baselines. AUC is shown with median and 50% and 95% HPD intervals.

Figure 5.4 Results of the unknown person detection experiment on test images

to some training data, and a flatter 'anomalous' distribution (Chandola et al., 2009, Sec. 7.1.3).[32]

We selected the 19 celebrities with at least 40 pictures available in LFW and randomly split them in two groups: 10 known and 9 unknown people. We used 27 images of each of the *known* people as training data and a disjoint test set of 13 images of each of the *known* and *unknown* people. We therefore have a binary classification setting with well-balanced classes at test time. Here, we ran our Gibbs sampler for 500 steps, discarding the first 100 burn-in iterations and thinning by a factor of 10, resulting in 320 pooled samples.

In Fig. 5.4a, we visualise the agreements between maximum *a posteriori* (MAP) identity predictions for test images:

$$\hat{z}_n = \arg\max_i p(z_n = i \,|\, \mathbf{x}_n, \mathbf{z}, \mathbf{c}^*, \boldsymbol{\pi}_0, \boldsymbol{\theta}^*), \tag{5.18}$$

where i ranges from 1 to $I+1$, the latter indicating an *unknown* identity, absent from the training set, and n indexes the test instances. Despite occasional ambiguous cases, the proposed model seems able to consistently group together all unknown faces, while successfully distinguishing between known identities.

As a simple baseline detector for comparison, we consider a threshold on the distance to the nearest neighbour (NN) in the face feature space (Chandola et al., 2009, Sec. 5.1). We also evaluate the decision function of a one-class SVM (Schölkopf et al., 2001), using an RBF kernel with $\gamma = 10$, chosen via leave-one-person-out cross-validation on the training set (roughly equivalent to thresholding the training data's kernel density estimate with bandwidth $1/\sqrt{2\gamma} \approx 0.22$). We compare the effectiveness

[32] The predictive distribution of $\mathbf{x}_n$ for new identities is a wide Student's t.

of both detection approaches using ROC curve analysis.

Figure 5.4b shows that, while all methods are highly effective at detecting unknown faces, scoring 95%+ AUC, ours consistently outperforms, by a small margin, both the NN baseline and the purpose-designed one-class SVM. Taking the MAP prediction, our model achieves [92.3%, 94.3%] detection accuracy.

5.5.2 Experiment 2: Identity discovery

We then investigate the clustering properties of the model in a purely unsupervised setting, when only context is provided. We evaluate the consistency of the estimated partitions of images into identities with the ground truth in terms of the adjusted Rand index (Hubert and Arabie, 1985; Rand, 1971).

Using simulations, besides having an endless source of data with ground-truth context and identity labels, we have full control over several important aspects of experimental setup, such as sequence lengths, rates of encounters, numbers of distinct contexts and people and amount of provided labels. Below we describe the simulation algorithm used in our experiments and illustrated in Fig. 5.5.

In our experiments we aim to simulate two important aspects of real-world identity recognition settings: 1. *Context*: knowing the context (e.g. location or time) makes it more likely for us to observe a particular subset of people; and 2. *Temporal consistency*: identities will not appear and disappear at random but instead be present for a longer duration.

To reproduce contexts, we simulate a single session of a user meeting new people. To this end we first create a number of fixed contexts and then assign identities uniformly at random to each context. For these experiments, we defined three contexts: 'home', 'work' and 'gym'. At any time, the user knows its own context and over time transitions between contexts. Independently at each frame, the user may switch context with a small probability.

To simulate temporal consistency, each person in the current context enters and leaves the camera frame as an independent binary Markov chain. As shown in Fig. 5.5 this naturally produces grouped observations. The image that is observed for each 'detected' face is sampled from the person's pictures available in the database. We sample these images without replacement and in cycles, to avoid observing the same image consecutively.

For this set of experiments, we consider three practical scenarios:

- *Online:* data is processed on a frame-by-frame basis, i.e. we extend the training set after each frame and run the Gibbs sampler for 10 full iterations
- *Batch:* same as above, but enqueue data for 20 frames before extending the training set and updating the model for 200 steps

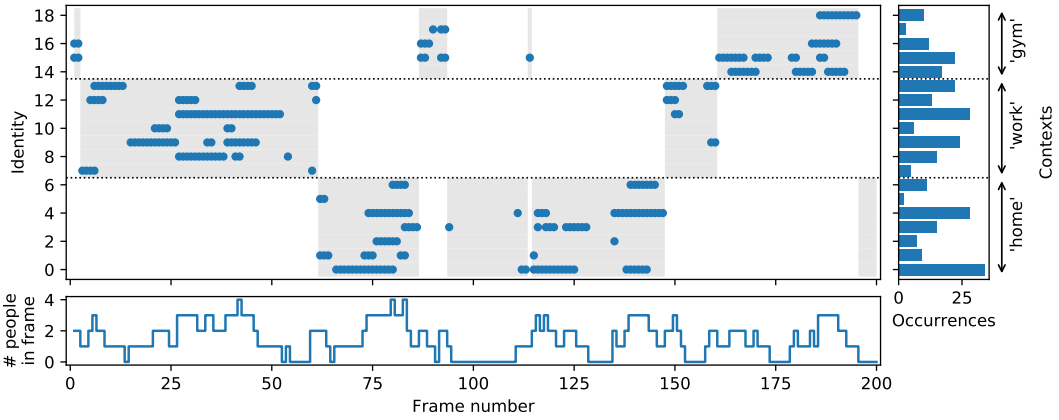
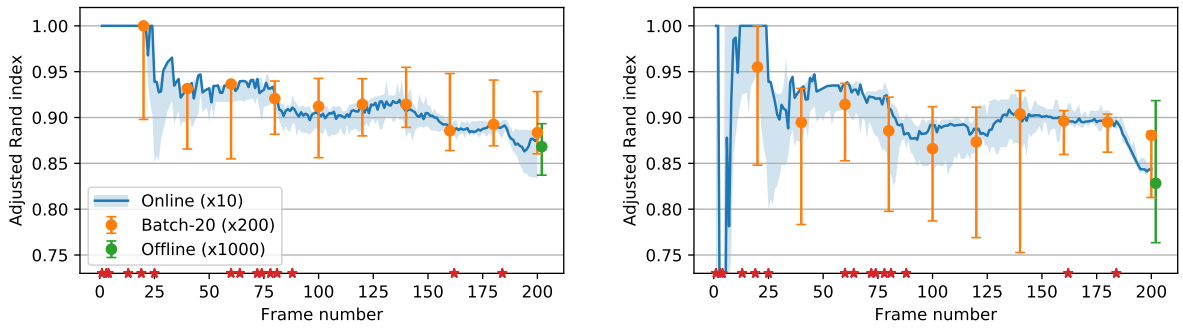


Figure 5.5 The simulation used in Experiment 2, showing identities coming in and out of the camera frame. Identities are shown grouped by their context (far right), and shading indicates identities present in the user's current context.



a With contexts **b** Without contexts

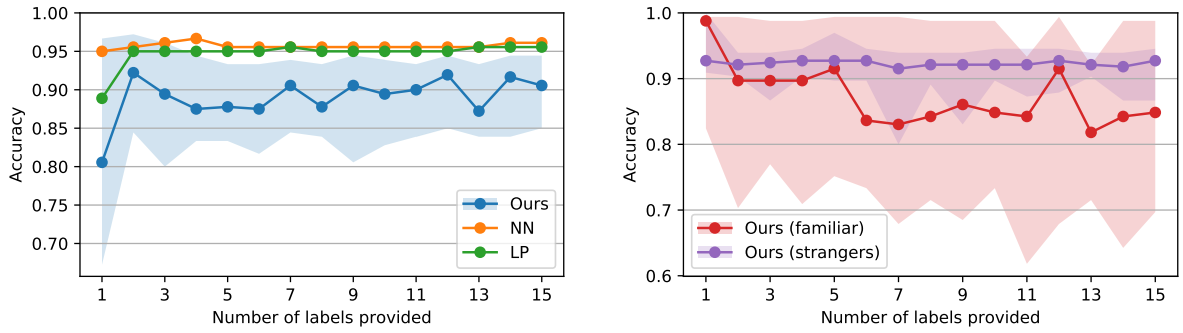
Figure 5.6 Identity clustering consistency. Markers on the horizontal axis (★) indicate when new people are met for the first time.

- *Offline:* assume entire sequence is available at once and iterate for 1000 steps

In the interest of fairness, the number of steps for each protocol was selected to give them roughly the same overall computation budget (ca. 200 000 frame-wise steps). In addition, we also study the impact on recognition performance of disabling the context model, by setting $C = 1$ and $c^*_m = 1, \forall m$.

We show the results of this experiment in Fig. 5.6. Clearly it is expected that, as more identities are met over time, the problem grows more challenging and clustering performance tends to decrease. Another general observation is that online processing produced much lower variance than batch or offline in both cases. The incremental availability of training data therefore seems to lead to more coherent states of the model.

Now, comparing Figs. 5.6a and 5.6b, it is evident that context-awareness not only reduces variance but also shows marginal improvements over the context-oblivious variant. Thus, without hurting recognition performance, the addition of a context model enables the *prediction* of context at test time, which may be useful for downstream user-experience systems.



a Acquaintances **b** Familiar and strangers

Figure 5.7 Label prediction accuracy. Note that NN and LP effectively have null accuracy for the *familiar* and *strangers* groups, as they cannot predict 'unknown'.

5.5.3 Experiment 3: Semi-supervised labelling

In our final set of experiments, we aimed to validate the application of the proposed label model for semi-supervised learning with sparse labels.

In the context of face identification, we may define three groups of people:

- *Acquainted:* known identity with known name
- *Familiar:* known identity with unknown name
- *Stranger:* unknown identity

We thus selected the 34 LFW celebrities with more than 30 pictures, and split them roughly equally in these three categories at random. From the *acquainted* and *familiar* groups, we randomly picked 15 of their images for training and 15 for testing, and we used 15 pictures of each *stranger* at test time only. We evaluated the label prediction accuracy as we varied the number of labelled training images provided for each acquaintance, from 1 to 15.

For baseline comparison, we evaluate nearest-neighbour classification (NN) and label propagation (LP; Zhu and Ghahramani, 2002), a similarity graph-based semi-supervised algorithm. We computed the LP edge weights with the same kernel as the SVM in Section 5.5.1. Recall that the face embedding network was trained with a triplet loss to explicitly optimise Euclidean distances for classification (Amos et al., 2016). As both NN and LP are distance-based, they are therefore expected to hold an advantage over our model for classifying labelled identities.

Section 5.5.3 shows the label prediction results for the labelled identities (acquaintances). In this setting, NN and LP performed nearly identically and slightly better than ours, likely due to the favourable embedding structure. Moreover, all methods predictably become more accurate as more supervision is introduced in the training data.

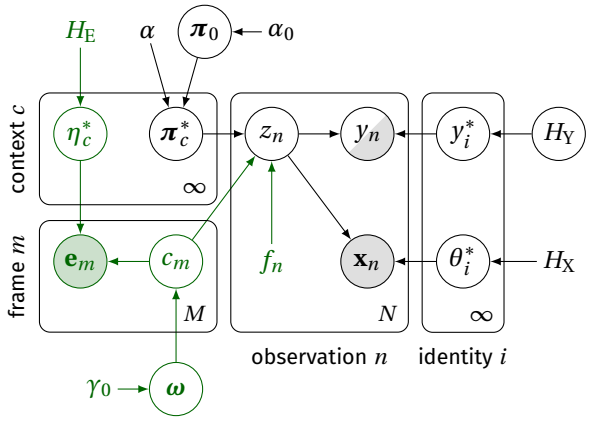


Figure 5.8 Overview of the extended probabilistic model. In addition to the variables in Fig. 5.2, this model includes context-dependent environmental cues, $\mathbf{e}_m$, and assumes an unbounded number of contexts.

More importantly, the key distinctive capabilities of our model are demonstrated in Section 5.5.3. As already discussed in Section 5.5.1, the proposed model is able to detect complete strangers, and here we see that it correctly predicts that their name is unknown. Furthermore, our model can acknowledge that familiar faces belong to different people, whose names may not be known. Neither of these functionalities is provided by the baselines, as they are limited to the closed-set identification task.

5.6 Extensions to the context model

In the previous sections, we proposed a unified probabilistic model that coherently reasons about the observed images, identities, partial knowledge about names, and the situational context of each observation. Our nonparametric identity model effectively represents an unbounded population of identities, while taking contextual co-occurrence relations and sparse noisy labels into account.

The context framework employed so far assumes a finite collection of pre-specified contexts and is fully supervised: an explicit context label c_m is observed with each frame m. This simplified scenario was adopted as a proof-of-concept, yet is admittedly unrealistic. In this section, we extend the referred context model in two ways: we explore its limit with an unbounded number of contexts, uncovering a rich nonparametric structure, and we lay the foundations for incorporating environmental cues—such as timestamps and geographical locations of frames—in our model to improve unsupervised context discovery and prediction.

5.6.1 Unbounded contexts

As explained in Section 5.3.1, the original context model had a *finite* $\mathrm{Dir}(\frac{\gamma_0}{C},\ldots,\frac{\gamma_0}{C})$ prior. This is a strong limitation, as the number of contexts, C, has to be chosen in advance and cannot easily be changed on-the-fly as the user varies their routine. A natural extension of such model is to take its

limit as $C \to \infty$, while tying the values of all context-wise concentration hyperparameters ($\alpha_c = \alpha, \forall c$). Remarkably, this results in yet another Dirichlet process (Neal, 2000). In particular, up to a reordering of the contexts, the prior on context proportions, $\boldsymbol{\omega}$, becomes GEM(γ_0).

This transformation has interesting theoretical and practical implications: the resulting structure is a *nested-hierarchical Dirichlet process*.[33] As before, at the top level we have the global identity distribution, G_0, over face parameters and labels, and the context-specific identity distributions, $(G_c^*)_{c=1}^C$, follow a DP with G_0 as a base measure:

$$G_0 \mid H_Y \sim \mathrm{DP}(\alpha_0, H_X \otimes H_Y) \tag{5.19}$$

$$G_c^* \mid G_0 \sim \mathrm{DP}(\alpha, G_0), \qquad c = 1, \ldots, \infty, \tag{5.20}$$

a prototypical example of a hierarchical DP (HDP; Teh et al., 2006). Recall that the exposition in previous sections considered an explicit representation of the nonparametric model in terms of weights (i.e. global and contextual identity probabilities), $\boldsymbol{\pi}_0$ and $(\boldsymbol{\pi}_c)_{c=1}^C$, and atom locations (i.e. identity-specific face parameters and true name labels), $(\theta_i^*, y_i^*)_{i=1}^\infty$. We can then relate the two representations as follows:

$$G_0 = \sum_{i=1}^{\infty} \pi_{0i} \delta_{(\theta_i^*, y_i^*)} \qquad \text{and} \qquad G_c^* = \sum_{i=1}^{\infty} \pi_{ci} \delta_{(\theta_i^*, y_i^*)}, \tag{5.21}$$

where $(\theta_i^*, y_i^*) \sim H_X \otimes H_Y$, for $i = 1, \ldots, \infty$.

Now, the nonparametric distribution of contexts implies wrapping the bottom level of the HDP, Eq. (5.20), as base for another DP, to form a nested DP (Blei et al., 2010; Rodríguez et al., 2008):

$$Q \mid G_0 \sim \mathrm{DP}(\gamma_0, \mathrm{DP}(\alpha, G_0)) \tag{5.22}$$

$$G_m \mid Q \sim Q = \textstyle\sum_{c=1}^{\infty} \omega_c \delta_{G_c^*}, \qquad m = 1, \ldots, M. \tag{5.23}$$

This construction inherits desirable properties from both elements: the hierarchy ensures that all frame-wise identity distributions, $(G_m)_{m=1}^M$, have the same support, and nesting produces clusters of frames with shared identity weights (i.e. contexts).

5.6.2 Environmental cues

While a purely identity-driven unsupervised context model may be able to disentangle co-occurrence patterns given enough data, we believe that environmental cues—such as timestamp and GPS coordin-

[33] This is related to the dual-HDP described in Wang et al. (2009) and the single-entity model of Agrawal et al. (2013), for example, although these works tended to focus on textual topic modelling.

ates of an acquired frame, if available—could considerably facilitate context discovery and prediction, in turn improving inference about identities.

Let us define $\mathbf{e}_m$ as the environmental measurements available for frame m, F_{E} a likelihood family parametrised by η_m, and H_{E} a prior distribution for such parameters. Plugging $\mathrm{DP}(\alpha, G_0) \otimes H_{\mathrm{E}}$ as base measure for the nested DP Q in Eq. (5.22), we can write

$$\eta_c^* \sim H_{\mathrm{E}}, \qquad c = 1, \ldots, \infty \tag{5.24}$$

$$\mathbf{e}_m \,|\, c_m, \boldsymbol{\eta}^* \sim F_{\mathrm{E}}(\eta_{c_m}^*), \qquad m = 1, \ldots, M. \tag{5.25}$$

Next, we present some preliminary ideas for defining a spatiotemporal context model.

One may define a spatial distribution centred on a geographic location where the specific context is most likely to occur, with a length scale in the order of e.g. tens of metres. At this range, the Earth's curvature is negligible and it may be possible to use a two-dimensional Gaussian distribution over latitude and longitude, as an approximation to a more correct (but more computationally expensive) spherical distribution (Mardia and Jupp, 1999), such as a tangential Gaussian (Straub et al., 2015).

We stipulate that a temporal model would have to accommodate both recurring and occasional contexts. One way to formalise such a model is as a Cox process (Cox, 1955), i.e. a Poisson process whose intensity function is itself random. For example, for recurring contexts (e.g. working hours, regular activities) we may consider the user's periodic weekly schedule divided into hourly bins. These can co-exist with one-off contexts (e.g. parties, meetings), which are localised in time. The resulting model is a mixture of context-specific Cox processes, which gives a certain probability that the user is in each context for each hour of the day.

5.7 Conclusion

In this work, we introduced a fully Bayesian treatment of the face identification problem. Each component of our proposed approach was motivated from human intuition about face recognition and tagging in daily social interactions. Our principled identity model can contemplate an unbounded population of identities, accounting for context-specific probabilities of meeting them.

We demonstrated that the proposed identity model can accurately detect when a face is unfamiliar, and is able to incrementally learn to differentiate between new people as they are met in a streaming data scenario. Lastly, we verified that our approach to dealing with sparse name annotations can handle not only acquaintances, whose names are known, but also familiar faces and complete strangers in a unified manner—a functionality unavailable in conventional (semi-) supervised identi-

fication methods.

We further proposed a nonparametric extension of the context model enabling unbounded context discovery, and discussed some of its theoretical implications in terms of nested-hierarchical nonparametric structures. Finally, we briefly examined how available environmental cues could be integrated into the model to replace the simplified, fully supervised context structure.

Here we explored the specific application domain of face recognition, analysing functionalities such as outlier detection, identity discovery, and semi-supervised tagging. When developing new models for heterogeneous data, there are multiple different aspects one may wish to evaluate: prediction, generation, reconstruction, imputation, abstraction, robustness, sensitivity, interpretability, etc. Unfortunately, it is difficult to find suitable testbeds for such comprehensive validation. We argue that an ideal testbed should be small and self-contained, with low computational requirements for quick development iterations, and also be flexible enough to allow validating all capabilities of interest. In addition—and most importantly—it ought to provide controllable experimental settings under which the expected model behaviour is well understood. Addressing these desiderata, the next chapter introduces a toolkit that transforms the popular MNIST dataset (LeCun et al., 1998) into a fully featured, multi-purpose experimental playground.

Chapter 6

Morpho-MNIST: A sandbox for learning with heterogeneous data

This chapter is based on the following publication:

- **Daniel C. Castro**, Jeremy Tan, Bernhard Kainz, Ender Konukoglu and Ben Glocker (2019). 'Morpho-MNIST: quantitative assessment and diagnostics for representation learning'. *Journal of Machine Learning Research* 20, 178.

Data and code are available at https://github.com/dccastro/Morpho-MNIST.

Revealing latent structure in data is an active field of research, having introduced exciting technologies such as variational autoencoders and adversarial networks, and is essential to push machine learning towards unsupervised knowledge discovery. However, a major challenge is the lack of suitable benchmarks for an objective and quantitative evaluation of learned representations. To address this issue we introduce Morpho-MNIST, a framework that aims to answer: "to what extent has my model learned to represent specific factors of variation in the data?" We extend the popular MNIST dataset by adding a morphometric analysis enabling quantitative comparison of trained models, identification of the roles of latent variables, and characterisation of sample diversity. We further propose a set of quantifiable perturbations to assess the performance of unsupervised and supervised methods on challenging tasks such as outlier detection and domain adaptation.

6.1 Introduction

A key factor for progress in machine learning has been the availability of well curated, easy-to-use, standardised and sufficiently large annotated datasets for benchmarking different algorithms and

models. This has led to major advances in speech recognition, computer vision, and natural language processing. A commonality between these tasks is their natural formulation as supervised learning tasks, wherein performance can be measured in terms of accuracy on a test set.

The general problem of representation learning (i.e. to reveal latent structure in data) is more difficult to assess due the lack of suitable benchmarks. Although the field is very active, with many recently proposed techniques such as probabilistic autoencoders and adversarial learning, it is less clear where the field stands in terms of progress or which approaches are more expressive for specific tasks. The lack of reproducible ways to quantify performance has led to subjective means of evaluation: visualisation techniques have been used to show low-dimensional projections of the latent space and visual inspection of generated or reconstructed samples are popular to provide subjective measures of descriptiveness. On the other hand, the quality of sampled images generally tells us little about how well the learned representations capture known factors of variation in the training distribution. In order to advance progress, the availability of tools for objective assessment of representation learning methods seems essential yet lacking.

This paper introduces Morpho-MNIST, a collection of shape metrics and perturbations, in a step towards quantitative assessment of representation learning. We build upon one of the most popular machine learning benchmarks, MNIST, which despite its shortcomings remains widely used. While MNIST was originally constructed to facilitate research in image classification, in the form of recognising handwritten digits (LeCun et al., 1998), it has found its use in representation learning, for example, to demonstrate that the learned latent space yields clusters consistent with digit labels. Methods aiming to disentangle the latent space claim success if individual latent variables capture specific style variations (e.g. stroke thickness, sidewards leaning digits and other visual characteristics).

It is important to emphasise that MNIST is a prime representative for datasets of rasterised two-dimensional shapes. In this type of data, while the underlying shapes may have relatively simple and well understood factors of variation, they manifest in the raster images as high-order correlations between large numbers of pixels. The challenge for a learning agent is to abstract away the low-level pixel intensity variations and recover a meaningful shape representation. MNIST is one of the interesting cases wherein it is possible to explicitly measure morphological attributes from the images themselves, enabling us to directly evaluate to what extent a trained model has learned to represent them.

More generally, MNIST remains popular in practice due to several factors: it allows reproducible comparisons with previous results reported in the literature; the dataset is sufficiently large for its complexity and consists of small, two-dimensional greyscale images defining a tractable ten-class classification problem; computation and memory requirements are low; most popular deep

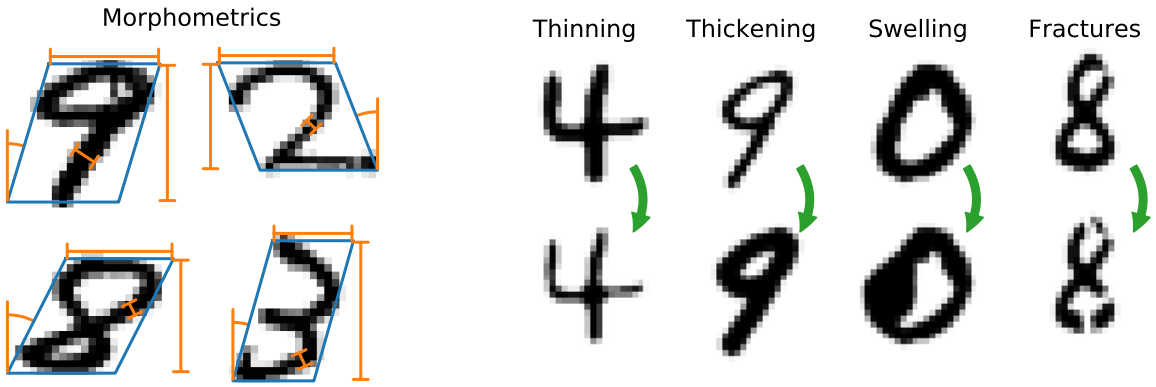


Figure 6.1 Overview of Morpho-MNIST. *Left:* MNIST morphometrics—stroke thickness and length (not shown), width, height and slant of digits. *Right:* MNIST perturbations (many more examples of each type in Appendix B.2).

learning frameworks and libraries offer tutorials using MNIST, which makes it straightforward for new researchers to enter the field and to experiment with new ideas and explore latest developments. Furthermore, there is now renewed interest in MNIST due to the recent replication of the original generation pipeline and the rediscovery of additional 50,000 test images, beckoning researchers to re-evaluate over 20 years of research on MNIST (Yadav and Bottou, 2019). We leverage these qualities and extend MNIST in multiple ways, as summarised in the following.

6.1.1 Contributions

Our aim is to provide a useful resource to the machine learning community, opening a variety of new avenues for experimentation and analysis. As we preserve the general properties of MNIST—such as image size, file format, numbers of training and test images, and the original ten-class classification problem—we hope this new quantitative framework for assessing representation learning will experience widespread adoption and may inspire further extensions facilitated by a publicly available Morpho-MNIST code base.

Morphometrics We propose to describe real and generated digit images in terms of measurable shape attributes. These include stroke thickness and length, and the width, height, and slant of digits (see Fig. 6.1, left). Whereas some of these MNIST properties have been analysed only visually in previous work, we demonstrate that quantifying each of them allows us to objectively characterise the roles of inferred representations.

More generally, the exclusive reliance on synthetic annotated datasets (Section 6.1.2) has been a crucial limitation for research into evaluation of learned representations (Eastwood and Williams, 2018), as this restricts us to the analysis of *inferential* behaviour (data → representation), even when

studying generative models. The tools we introduce here can be used to measure model samples, effectively enabling direct assessment of *generative* performance (representation → data).

These measurements can be directly employed to re-evaluate existing models and may be added retrospectively to previous experiments involving the original MNIST dataset. Adoption of our morphometric analysis may provide new insights into the effectiveness of representation learning methods in terms of revealing meaningful latent structures. Furthermore, for other datasets it suffices to design the relevant scalar metrics and include them in the very same evaluation framework.

Perturbations We introduce a set of parametrisable global and local perturbations, inspired by natural and pathological variability in medical images. Global changes involve overall thinning and thickening of digits, while local changes include both swelling and fractures (see examples on the right in Fig. 6.1 and many more in Appendix B.2). The perturbation framework we present in Section 6.3—illustrating various ways to exploit the outputs of the proposed processing pipeline (e.g. skeleton, distance map, and even morphometrics themselves; see Section 6.2.1)—is sufficiently general that additional transformations can easily be designed as necessary. Injecting these perturbations into the dataset adds a new type of complexity to the data manifold and opens up a variety of interesting applications.

The proposed perturbations are designed to enable a wide range of new studies and applications for both supervised and unsupervised tasks. Detection of 'abnormalities' (i.e. local perturbations) is an evident application, although more challenging tasks can also be defined, such as classification from noisy/corrupted data, domain adaptation, localisation of perturbations, characterising semantics of learned latent representations, and more. We explore a few supplementary examples of supervised tasks in Section 6.4.3.

6.1.2 Related work: Datasets

In this section, we provide an overview of some datasets that are related to MNIST, by either sharing its original source content, containing transformations of the original MNIST images or being distributed in the same format for easy replacement. We also mention a few prevalent datasets of synthetic images with generative factor annotations, analogous to the morphometrics proposed in this paper.

NIST datasets The MNIST (modified NIST) dataset (LeCun et al., 1998) was constructed from handwritten digits in NIST Special Databases 1 and 3, now released as Special Database 19 (Grother and Hanaoka, 2016). G. Cohen et al. (2017) generated a much larger dataset based on the same NIST

database, containing additional upper- and lower-case letters, called EMNIST (extended MNIST). LeCun et al.'s original processing pipeline was recently replicated, enabling the rediscovery of 50,000 test images that had been excluded from MNIST—a dataset now released as QMNIST (Yadav and Bottou, 2019).

MNIST perturbations The seminal paper by LeCun et al. (1998) employed data augmentation using planar affine transformations including translation, scaling, squeezing, and shearing. Loosli et al. (2007) employed random elastic deformations to construct the Infinite MNIST dataset. Other MNIST variations include rotations and insertion of random and structured background (Larochelle et al., 2007), and Tieleman (2013) applied spatial affine transformations and provided ground-truth transformation parameters.

MNIST format Due to the ubiquity of the MNIST dataset in machine learning research and the resulting multitude of compatible model architectures available, it is appealing to release new datasets in the same format (28×28, 8-bit grayscale images). One such effort is Fashion-MNIST (Xiao et al., 2017), containing images of clothing articles from ten distinct classes, adapted from an online shopping catalogue. Another example is notMNIST (Bulatov, 2011), a dataset of character glyphs for letters 'A'–'J' (also ten classes), in a challengingly diverse collection of typefaces.

Annotated datasets Synthetic computer vision datasets that are popular for evaluating disentanglement of learned latent factors of variation include those from Paysan et al. (2009) and Aubry et al. (2014). They contain 2D renderings of 3D faces and chairs, respectively, with ground-truth pose parameters (azimuth, elevation) and lighting conditions (faces only). A further initiative in that direction is the dSprites dataset (Matthey et al., 2017), which consists of binary images containing three types of shapes with varying location, orientation and size. The availability of the ground-truth values of such attributes has motivated the accelerated adoption of these datasets in the evaluation of representation learning algorithms.

6.1.3 Related work: Quantitative evaluation

Evaluation of representation learning is a notoriously challenging task and remains an open research problem. Numerous solutions have been proposed, with many of the earlier ones focusing on the test log-likelihood under the model (Kingma and Welling, 2014a) or, for likelihood-free models, under a

kernel density estimate (KDE) of generated samples (Goodfellow et al., 2014; Makhzani et al., 2016)—shown not to be reliable proxies for the true model likelihood (Theis et al., 2016).

Another perspective for evaluation of generative models of images is the visual fidelity of its samples to the training data, which would normally require manual inspection. To address this issue, a successful family of metrics have been proposed, based on visual features extracted by the Inception network (Szegedy et al., 2016). The original Inception score (Salimans et al., 2016) relies on the 'crispness' of class predictions, whereas the Fréchet Inception distance (FID; Heusel et al., 2017) and the kernel Inception distance (KID; Bińkowski et al., 2018) statistically compare high-level representations instead of the final network outputs.

Although the approaches above can reveal vague signs of mode collapse, it may be useful to diagnose this phenomenon on its own. With this objective, Arora et al. (2018) proposed to estimate the support of the learned distribution (assumed discrete) using the birthday paradox test, by counting pairs of visual duplicates among model samples. Unfortunately, the adoption of this technique is hindered by its reliance on manual visual inspection to flag identical images.

There have been several attempts at quantifying representation disentanglement performance. For example, Higgins et al. (2017) proposed to use the accuracy in predicting which factor of variation was held fixed in a simulated dataset. There exist further information-theoretic approaches, involving the KL divergence contribution from each latent dimension (Dupont, 2018) or their mutual information with each known generative factor (R. T. Q. Chen et al., 2018a). Yet another method, explored in Kumar et al. (2018), is based on the predictive accuracy of individual latent variables to each generative factor. The concurrent work of Eastwood and Williams (2018) introduced a comprehensive methodology for characterising different aspects of representation disentanglement.

6.2 Morphometry

Meaningful morphometrics are instrumental in characterising distributions of rasterised shapes, such as MNIST digits, and can be useful as additional data for downstream learning tasks. We regard the current choice of morphometrics as a minimal useful collection of attributes that are largely independent and can be estimated robustly. Although further more intricate metrics can be defined, they are likely to be entangled with the ones defined here and to be noisy due to the limited resolution of the original images. We begin this section by describing the image processing pipeline employed for extracting the metrics and for applying perturbations (Section 6.3), followed by details on the computation of each measurement.

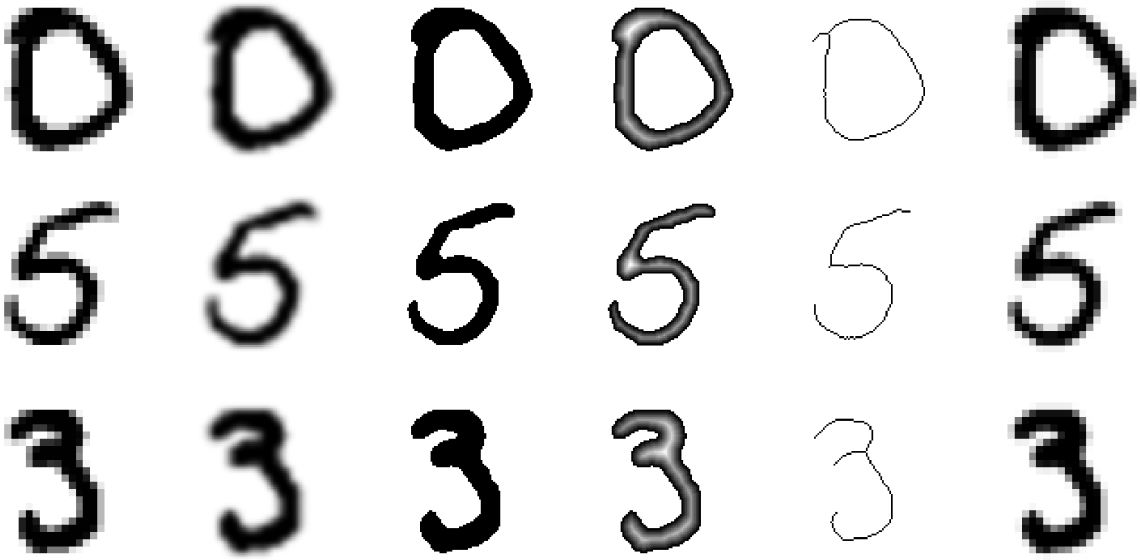


Figure 6.2 Stages of the image processing pipeline. *Left to right:* original image, upscaled image, binarised image, distance transform, skeleton, downscaled image.

6.2.1 Processing pipeline

The original 28×28 resolution of the MNIST images is generally not high enough to enable satisfactory morphological processing: stroke properties (e.g. length, thickness) measured directly on the binarised images would likely be inaccurate and heavily quantised. To mitigate this issue and enable sub-pixel accuracy in the measurements, we propose to use the following processing steps:

1. *Upscale* (e.g. ×4, to 112×112);[34]
2. *Binarise:* the blurry upscaled image is thresholded at half its intensity range (e.g. 128 for most images), which ensures thin or faint digits are not erased;
3. Compute *Euclidean distance transform* (EDT): each pixel within the digit boundaries contains the distance to its nearest boundary point;
4. *Skeletonise:* detect the ridges of the EDT, i.e. the locus of points equidistant to two or more boundary points (also known as the *medial axis*; H. Blum, 1967);
5. Optionally apply *perturbation* to binarised image (cf. Section 6.3);
6. *Downscale* binarised or perturbed image to original resolution.

We illustrate the pipeline in Fig. 6.2. The binary high-resolution digits have smooth boundaries and faithfully capture subtle variations in contour shape and stroke thickness that are only vaguely

[34] Up- and downscaling by a factor of f are done with bicubic interpolation and Gaussian smoothing (bandwidth $\sigma = 2f/6$), following `scikit-image` defaults (Walt et al., 2014).

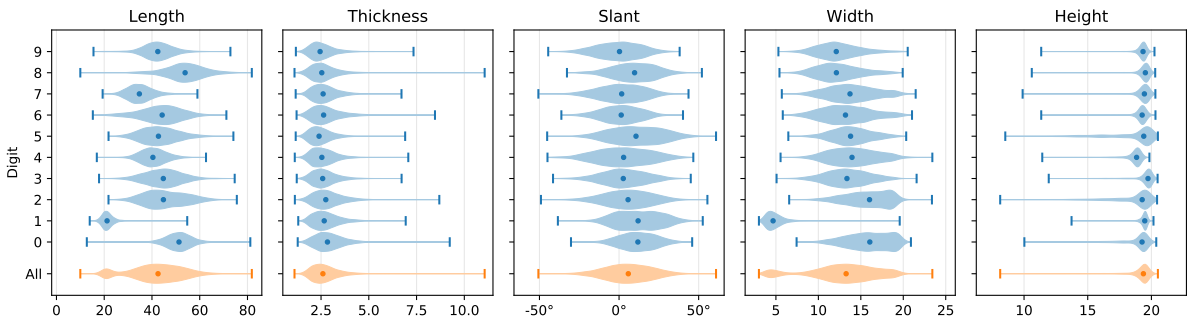


Figure 6.3 Distribution of morphological attributes in the plain MNIST training dataset

discernible in the low-resolution images. Additionally, note how the final downscaled image is almost indistinguishable from the original.

All morphometric attributes described below are calculated for each digit after applying steps 1–4 of this pipeline. The distributions for plain MNIST are plotted in Fig. 6.3, and the distributions after applying each type of perturbation can be found in Appendix B.1.

6.2.2 Stroke length

Here we approximate the trace of the pen tip, as a digit was being written, by the computed morphological skeleton. In this light, the total length of the skeleton is an estimate of the length of the pen stroke, which in turn is a measure of shape complexity.

It can be computed in a single pass by accumulating the Euclidean distance of each skeleton pixel to its immediate neighbours, taking care to only count the individual contributions once. This approach is more accurate and more robust against rotations than a naïve estimate by simply counting the pixels.

6.2.3 Stroke thickness

A prominent factor of style variation in the MNIST digits is the overall thickness of the strokes, due to both legitimate differences in pen thickness and force applied, and also to the rescaling of the original NIST images by different factors.

We estimate it by exploiting both the computed distance transform and the skeleton. Since skeleton pixels are equidistant to the nearest boundaries and the value of the EDT at those locations corresponds to the local half-width (H. Blum, 1967), we take twice the mean of the EDT over all skeleton pixels as our global estimate for stroke thickness.

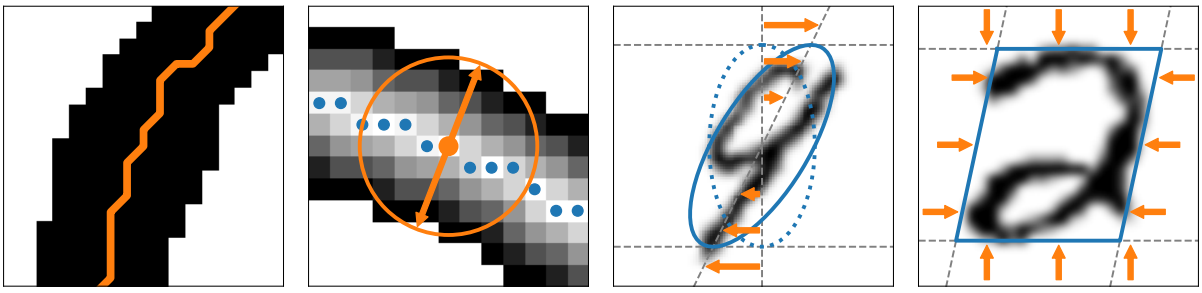


Figure 6.4 Measuring shape attributes. *Left to right:* stroke length (length of binarised image's skeleton); stroke thickness (average distance transform over skeleton pixels ×2); slant (horizontal shearing angle); width and height (bounding parallelogram).

6.2.4 Slant

The extent by which handwritten symbols lean right or left (forward and backward slant, respectively) is a further notorious and quantifiable dimension of handwriting style. It introduces so much variation in the appearance of characters in images that it is common practice in optical character recognition (OCR) systems to 'deslant' them, in an attempt to reduce within-class variance (LeCun et al., 1998; Teow and Loe, 2002).

We adapt the referred deslanting methodology to describe the slant *angle* of the handwritten digits. After estimating the second-order image moments,

$$\mathbf{S} = \begin{pmatrix} \sum_{i,j} x_{ij}(i-\bar{\imath})^2 & \sum_{i,j} x_{ij}(i-\bar{\imath})(j-\bar{\jmath}) \\ \sum_{i,j} x_{ij}(i-\bar{\imath})(j-\bar{\jmath}) & \sum_{i,j} x_{ij}(j-\bar{\jmath})^2 \end{pmatrix}, \tag{6.1}$$

we define the slant based on the horizontal shear:

$$\alpha = \arctan\left(-\frac{S_{12}}{S_{22}}\right), \tag{6.2}$$

where x_{ij} is the intensity of the pixel at row j and column i, and $(\bar{\imath}, \bar{\jmath})$ are the centroid coordinates. The minus sign ensures that positive and negative values of α correspond to forward and backward slant, respectively.

Note that $-\alpha$ is the angle by which the image would need to be horizontally sheared toward the vertical centre line such that the digit is upright (i.e. principal axes aligned with the coordinate axes). Note how this differs from a *rotation* angle, which would instead be defined in terms of $\frac{1}{2}\arctan(2S_{12}/(S_{22}-S_{11}))$.

6.2.5 Width and height

It is useful to measure other general shape attributes, such as width, height, and aspect ratio, which also present substantial variation related to personal handwriting style.[35] To this end, we propose to fit a *bounding parallelogram* to each digit, with horizontal and slanted sides (see Fig. 6.1). We highlight that it does not suffice to simply fit a bounding box with perpendicular sides, since the width would clearly be confounded with slant. These metrics would therefore contain redundant information that obfuscates the true digit width.

The image is swept top-to-bottom with a horizontal boundary to compute a vertical marginal cumulative distribution function (CDF), and likewise left-to-right with a *slanted* boundary for a horizontal marginal CDF, with angle α as computed above. The bounds are then chosen based on equal-tailed intervals containing a given proportion of the image mass—98% in both directions (1% from each side) proved accurate and robust in our experiments.

6.3 Perturbations

As discussed in Section 6.1, we bring forward a number of morphological perturbations for MNIST digits, to enable diverse applications and experimentation. In this section, we detail these parametrisable transformations, categorised as global or local.

6.3.1 Global: Thinning and thickening

The first pair of transformations we present is based on simple morphological operations: the binarised image of a digit is dilated or eroded with a circular structuring element. Its radius is set proportionally to the estimated stroke thickness (Section 6.2.3), so that the overall thickness of each digit will decrease or increase by an approximately fixed factor (here, -70% and +100%; see Figs. B.6 and B.7).

Since there is substantial thickness variability in the original MNIST data (cf. Fig. 6.3) and most thinned and thickened digits look very plausible, we believe that these perturbations can constitute a powerful form of data augmentation for training. For the same reason, we have not included these perturbations in the abnormality detection experiments (Section 6.4.3.3).

[35] Little variation in height is expected, since the original handwritten digits were scaled to fit a 20×20 box (LeCun et al., 1998). Nevertheless, a minority of digits were originally wider than they were tall, which explains the long tails in the distribution of heights (Fig. 6.3).

6.3.2 Local: Swelling

In addition to the global transformations above, we introduce *local* perturbations with variable location and extent, which are harder to detect automatically. Given a radius R, a centre location $\mathbf{r}_0$ and a strength parameter $\gamma > 1$, the coordinates $\mathbf{r}$ of pixels within distance R of $\mathbf{r}_0$ are nonlinearly warped according to a radial power transform:

$$\mathbf{r} \mapsto \mathbf{r}_0 + (\mathbf{r} - \mathbf{r}_0)\left(\frac{\|\mathbf{r} - \mathbf{r}_0\|}{R}\right)^{\gamma-1}, \tag{6.3}$$

leaving the remaining portions of the image untouched and resampling with bicubic interpolation.

In the experiments and released dataset, we set $\gamma = 7$ and $R = 3\sqrt{\theta}/2$, where θ is thickness. Unlike simple linear scaling with θ, this choice for R produces noticeable but not exaggerated effects across the thickness range observed in the dataset (cf. Fig. B.8). The centre location, $\mathbf{r}_0$, is picked uniformly at random from the pixels along the estimated skeleton.

6.3.3 Local: Fractures

Finally, we describe the proposed procedure for adding fractures to an MNIST digit, where we define a fracture as a break in the continuity of a pen stroke. Because single fractures can in many cases be easily mistaken for true gaps between strokes, we add multiple fractures to each affected digit.

When selecting where to place a fracture, we attempt to avoid locations too close to stroke tips (points on the skeleton with a single neighbour) or fork points (more than two neighbours). This is achieved by sampling only among those skeleton pixels above a certain distance to these detected points. In addition, we desire fractures to be transversal to the pen strokes. Local orientation is determined based on second-order moments of the skeleton inside a window centred at the chosen location, and the length of the fracture is estimated from the boundary EDT. Finally, the fracture is drawn onto the high-resolution binary image with a circular brush along the estimated normal.

In practice, we found that adding three fractures with 1.5 px thickness, 2 px minimum distance to tips and forks and angle window of 5×5 px^2 ('px' as measured in the low resolution image) produces detectable but not too obvious perturbations (see Fig. B.9). We also extend the lines on both ends by 0.5 px to ensure some tolerance.

6.4 Evaluation case studies

In this section, we demonstrate potential uses of the proposed framework: using morphometrics to characterise the distribution of samples from generative models and finding associations between

learned latent representations and morphometric attributes. In addition, we exemplify in Section 6.4.3 a variety of novel ways to make use of morphometry and perturbations for evaluating predictive models on MNIST and beyond.

6.4.1 Sample diversity

Here we aim to illustrate ways in which the proposed MNIST morphometrics may be used to visualise distributions learned by generative models and to quantify their agreement with the true data distribution in terms of these concrete attributes. Moreover, extracting such measurements from model samples could be a step toward diagnosing mode collapse.

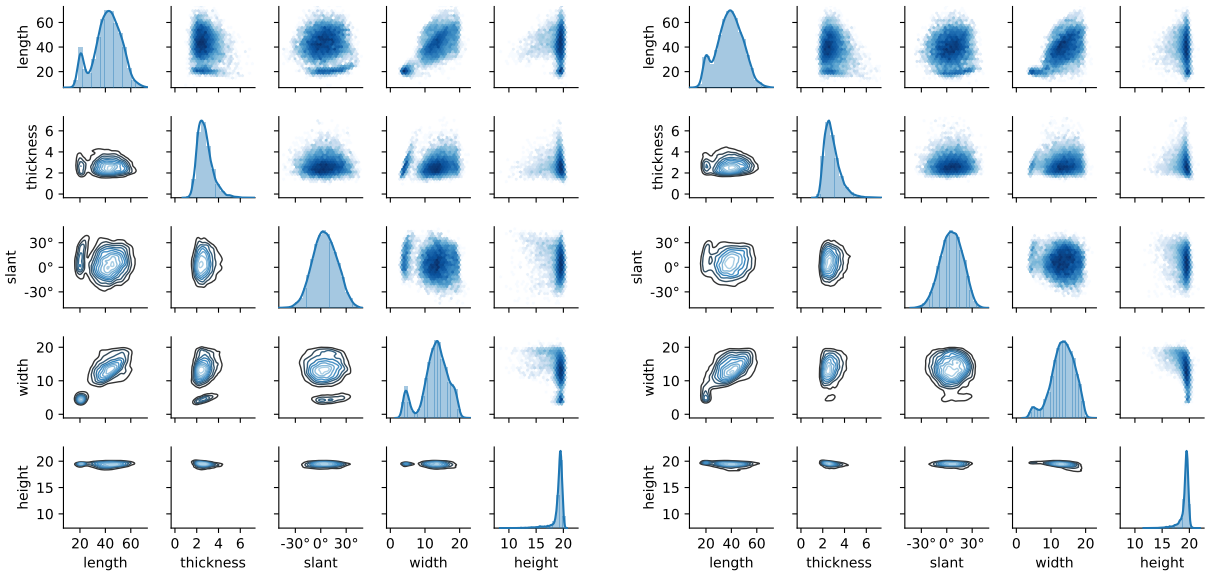
We exemplify this scenario with a generative adversarial network (GAN, Goodfellow et al., 2014) and a β-variational autoencoder (β-VAE, Higgins et al., 2017), both with generator (resp. decoder) and discriminator architecture as used in the MNIST experiments in X. Chen et al. (2016), and encoder mirroring the decoder. We train a β-VAE with $\beta = 4$, as in Higgins et al. (2017)'s experiments on small binary images, and a vanilla GAN with non-saturating loss, both with 64-dimensional latent space. To explore the behaviour of a much less expressive model, we additionally train a GAN with only two latent dimensions.

6.4.1.1 Visualisation

Figure 6.5 illustrates the morphometric distributions of the plain MNIST test images and of 10,000 samples from each of these three models. As can be seen, morphometrics provide *interpretable* low-dimensional statistics which allow comparing distributions learned by generative models between each other and with true datasets. While Figs. 6.5b and 6.5c show model samples roughly as diverse as the true images, the samples from the low-dimensional GAN in Fig. 6.5d seem concentrated on certain regions, covering a distribution that is less faithful to the true one in Fig. 6.5a.

6.4.1.2 Statistical comparison

We argue that in this lower-dimensional space of morphometrics it is possible to statistically compare the distributions, since this was shown not to be effective directly in image space (e.g. Theis et al., 2016). To this end, we propose to use kernel two-sample tests based on maximum mean discrepancy (MMD) between morphometrics of the test data and of each of the sample distributions (Gretton et al., 2012). The MMD statistic is a measure of dissimilarity between distributions, such that a value significantly greater than zero (small p-value) suggests a detectable difference. Here, we performed



a MNIST test images

b Samples from a β-VAE

c Samples from a GAN

d Samples from a low-dimensional GAN

Figure 6.5 Distribution of morphometric attributes for MNIST test dataset and samples from some generative models. Diagonals show marginal histograms and KDEs, upper-triangular plots show pairwise log-histograms and lower-triangular plots show pairwise KDEs.

Table 6.1 Kernel two-sample tests between model samples and true test data. The bold row indicates a significant divergence, thus a failure of the model in faithfully reproducing the data distribution.

Test data vs.	Dims.	$\mathrm{MMD}_l^2 \pm$ std. error ($\times 10^{-3}$)	p
β-VAE	64	0.792 ± 1.569	.3068
GAN	64	1.458 ± 1.650	.1885
GAN	2	$\mathbf{8.876 \pm 1.807}$	**.0000**

the linear-time asymptotic test described in Gretton et al. (2012, Sec. 6).[36]

The results of these tests are displayed in Table 6.1, where we observe that samples from the high-dimensional β-VAE and GAN did not exhibit a significantly large MMD, thus there is no evidence of distributional mismatch. On the other hand, the low-dimensional GAN's samples show a significant departure from the data distribution, confirming the qualitative judgement based on comparing Figs. 6.5a and 6.5d.

Whereas the bandwidth heuristic used here is fairly crude, much more sophisticated kernel selection procedures exist, e.g. by explicitly optimising the test power (Sutherland et al., 2017). A further analysis tool in a similar vein would be to apply a *relative* MMD similarity test (Bounliphone et al., 2016), to rank trained models based on sample fidelity. It would also be possible to adopt a model criticism methodology based on the MMD witness function (Lloyd and Ghahramani, 2015), to identify over- and under-represented regions in morphometric space (and corresponding generated image exemplars could also be inspected).

6.4.2 Disentanglement

In this experiment, we demonstrate that: (a) standard MNIST can be augmented with morphometric attributes to quantitatively study representations computed by an *inference* model (as already possible with annotated datasets, e.g. dSprites and 3D faces); (b) we can measure shape attributes of samples to assess disentanglement of a *generative* model, which is unprecedented to the best of our knowledge; and (c) this analysis can also diagnose when a model unexpectedly *fails* to disentangle a known aspect of the data.

As for the sample diversity experiments in Section 6.4.1, the purpose of this disentanglement analysis is to study the behaviour of specific trained model instances, and results may not generalise to the respective model classes or architectures, to different datasets, or even to distinct runs of the

[36] We employed a Gaussian product kernel with bandwidths derived from Scott's rule-of-thumb, analogously to the KDE plots in Fig. 6.5. This rule defines the bandwidth for a density estimation kernel as $N^{-1/(D+4)}$ times the standard deviation in each dimension, where N and D denote sample size and number of dimensions (Scott, 1992, Eq. 6.42). We determine the KDE bandwidths separately for real and sample data, then add their squares to obtain the squared bandwidth of the MMD's Gaussian kernel, as it corresponds to the *convolution* of the density estimation kernels chosen for each set of data. See Gretton et al. (2012, Sec. 3.3.1) for further details on the relation between MMD and L^2 distance of kernel density estimates.

same set-up. Similarly to any other evaluation methodology, findings concerning disentanglement patterns are reproducible only to the extent that the model under study can be reliably retrained to a similar state.

6.4.2.1 Methodology

Using an approach related to the disentanglement measure introduced by Kumar et al. (2018), we study the correlation structures between known generative factors and learned latent codes. Specifically, we compute the *partial correlation* ($r_{c_i y_j \cdot \mathbf{c}_{-i}}$) between each latent code variable (c_i) and each morphometric attribute (y_j), controlling for the variation in the remaining latent variables ($\mathbf{c}_{-i}$).[37] As opposed to the simple correlation, this technique allows us to study the *net* first-order effect of each latent code, all else being equal.

This partial correlation is calculated as follows, ranging over codes i and attributes j:

$$r_{c_i y_j \cdot \mathbf{c}_{-i}} = -\frac{P^{(j)}_{1,i+1}}{\sqrt{P^{(j)}_{11} P^{(j)}_{i+1,i+1}}} \quad \in [-1, 1]\,, \tag{6.4}$$

where $\mathbf{P}^{(j)}$ is the sample precision matrix (i.e. inverse of sample covariance) of $(y_j, c_1, \ldots, c_D)$ over all data points. Note that Eq. (6.4) is closely related to the least-squares coefficients for multiple linear regression of y_j on $(c_1, \ldots, c_D)$, and its magnitude can be directly interpreted as a relative importance score within the framework of Eastwood and Williams (2018).

We plot the calculated partial correlations in the form of tables, where each row corresponds to a shape attribute, and columns represent the latent variables. These tables should be read row-wise, as the scores are normalised separately for each shape factor. Entries close to zero do not indicate that the latent dimension is unrelated to the measured attribute, but rather that it is entangled with others and has little marginal effect on its own.

In addition, we demonstrate how our morphometry framework is fully compatible with contemporary disentanglement metrics by also computing the mutual information gap (MIG score; R. T. Q. Chen et al., 2018a) in each scenario. This score is based on the mutual information of each pair of generative factor (here, shape attribute) and latent code dimension, and reports the gap between the two leading codes for each factor, normalised to $[0, 1]$. It therefore quantifies how prominently the attribute is represented by a single latent dimension, i.e. its *completeness* in the terminology of Eastwood and Williams (2018). The overall MIG for each model is given by the average over all attributes.

[37] For the categorical code, c_1, we take a single binary dummy variable for each category, $c_1^{(k)}$, while controlling only for the remaining codes (c_2, c_3 etc.) to avoid multicollinearity.

Table 6.2 Settings for InfoGAN disentanglement experiments

	# Cat.	# Cont.	# Bin.	Dataset
InfoGAN-A	10	2	0	PLAIN: plain only
InfoGAN-B	10	3	0	GLOBAL: plain + thinning + thickening
InfoGAN-C	10	2	2	LOCAL: plain + swelling + fractures



a InfoGAN-A (overall MIG = 0.0582) **b** InfoGAN-B (overall MIG = 0.1748)

Figure 6.6 Partial correlations between inferred latent codes and morphometrics of test images. Circle area and colour strength are proportional to correlation magnitude, blue is positive and red is negative. On the right, we indicate the mutual information gap score (MIG) for each attribute.

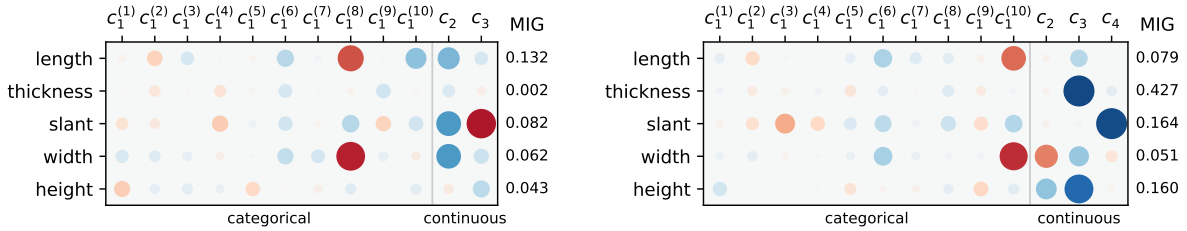
For these experiments, we trained InfoGAN (X. Chen et al., 2016) instances with various latent space configurations, then took MAP estimates of latent codes for each image (i.e. maximal logit for categorical codes and mean for continuous codes) as predicted by the variational recognition network. Models were trained for 20 epochs using 64 images per batch, with no hyperparameter tuning. We emphasise that our goal was to illustrate how the proposed morphometrics can serve as tools to better understand whether models behave as intended, and not to optimally train them in each scenario.

6.4.2.2 Inferential disentanglement

To illustrate how this methodology can be applied in practice to assess disentanglement, we consider two settings. The first is the same as in the MNIST experiment from X. Chen et al. (2016), with a 10-way categorical and two continuous latent codes, trained and evaluated on the plain MNIST digits, to which we will refer as InfoGAN-A.

The second setting was designed to investigate whether the model could disentangle the concept of *thickness*, by including an additional continuous latent code and training on a dataset with exaggerated thickness variations. We constructed this dataset by randomly interleaving plain, thinned and thickened digit images in equal proportions. Since the perturbations were applied completely at random, we expect a trained generative model to identify that thickness should be largely independent of the other morphological attributes. We refer to this set-up as InfoGAN-B. Table 6.2 summarises the different experimental settings, for reference.

In Fig. 6.6a, we see that InfoGAN-A learned to encode slant mostly in c_3, while $c_1^{(8)}$ clearly relates to the '1' class (much narrower digit shape and shorter pen stroke; cf. Fig. 6.3). Figure 6.6b



a InfoGAN-A (overall MIG = 0.0641) **b** InfoGAN-B (overall MIG = 0.1762)

Figure 6.7 Partial correlations between 1000 sampled latent codes and morphometrics of the corresponding generated images

quantitatively confirms the hypothesis that InfoGAN-B's recognition network would learn to separate slant and thickness (in c_4 and c_3, resp.), the most prominent factors of style variation in this dataset. Interestingly, it shows that c_3 also associates with height, as thicker digits also tend to be taller.

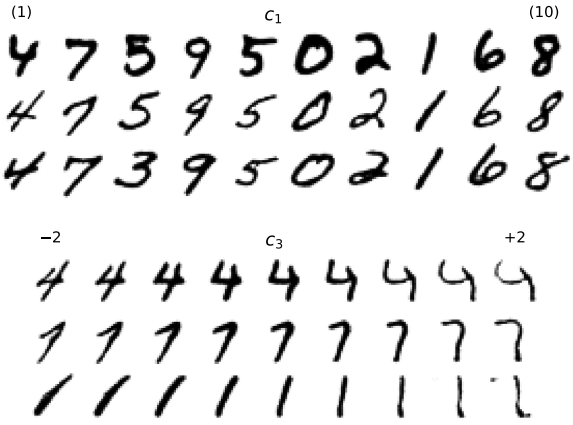
Moreover, we observe that the MIG for each attribute correlates with the qualitative interpretation of these partial correlation tables. For example, thickness is represented considerably more compactly in InfoGAN-B (mainly by c_3) than in InfoGAN-A (highly entangled), exhibiting a correspondingly much higher MIG score. For InfoGAN-B, a noticeable portion of the slant variation was additionally captured by c_1, which is reflected in its lower MIG than that for thickness. Lastly, the difference in overall MIG between models validates our qualitative judgement about global disentanglement.

6.4.2.3 Generative disentanglement

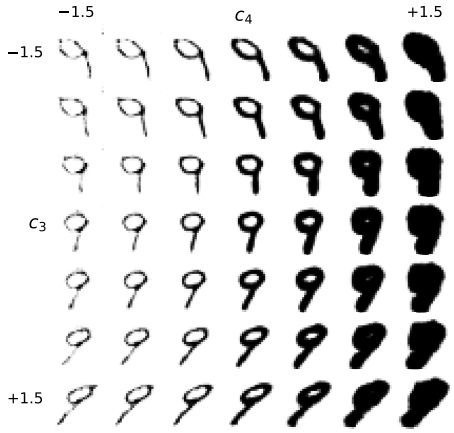
The evaluation methodology described above is useful for investigating the behaviour of the *inference* direction of a model, and can readily be used with datasets which include ground-truth generative factor annotations. On the other hand, unless we trust that the inference approximation is highly accurate, this tells us little about the *generative* expressiveness of the model. This is where computed metrics truly show their potential: one can measure generated samples, and investigate how their attributes relate to the latent variables used to create them.

Figure 6.7 shows results for a similar analysis to Fig. 6.6, but now evaluated on *samples* from that model. As the tables are mostly indistinguishable, we may argue that in this case the inference and generator networks have learned to consistently encode and decode the digit shape attributes. Further, the per-attribute and overall MIG scores agree once more with the observed disentanglement patterns.

As further illustration, Fig. 6.8 displays traversals of the latent space, obtained by varying a subset of the latent variables while holding the remaining ones (including noise) constant. With these examples, we can qualitatively verify the quantitative results in Fig. 6.7. Note that, until now, visual inspection was typically the only means of evaluating disentanglement and expressiveness of the generative direction of image models (H. Kim and Mnih, 2018).



a INFOGAN-A: one-dimensional traversals of c_1 (*top*, 'digit type') and c_3 (*bottom*, 'slant'). Samples in each row share the values of remaining latent variables and noise.



b INFOGAN-B: two-dimensional traversal of $c_4 \times c_3$ ('thickness' × 'slant')

Figure 6.8 InfoGAN latent space traversals

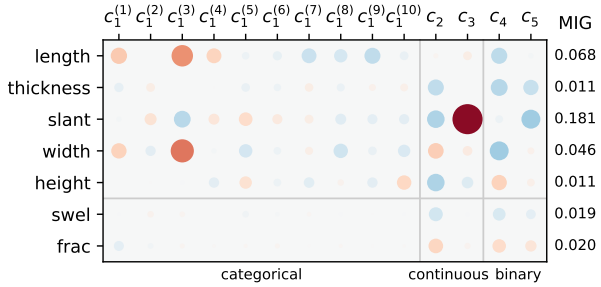


Figure 6.9 Partial correlations between inferred latent codes and morphometrics of test images for INFOGAN-C (overall MIG = 0.0507). 'swel' and 'frac' refer to the binary perturbation labels.

6.4.2.4 Diagnosing failure

We also attempted to detect whether an InfoGAN had learned to discover that local perturbations (swelling and fractures) correspond to independent generative factors. To this end, we extended the model formulation with additional Bernoulli latent codes, which may learn to encode presence/absence of each type of local perturbation. The model investigated here, dubbed INFOGAN-C (cf. Table 6.2), had a 10-way categorical, two continuous and two binary codes, and was trained with a dataset of plain, swollen and fractured digits (randomly mixed as above).

Again via inferential partial correlation analysis—now including ground-truth perturbation labels—we can quantitatively discern that this particular model instance was unable to meaningfully disentangle the perturbations (Fig. 6.9, bottom-right block). This is also confirmed by the low MIG for the corresponding rows. In fact, it appears that the addition of the binary variables did not lead to more expressive representations in this case, even impairing the disentanglement of the categorical variables, if compared to Figs. 6.6a and 6.6b, for example.

6.4.3 Predictive model evaluation

Although the driving motivation for introducing Morpho-MNIST has been the lack of means for quantitative evaluation of generative models, the proposed framework may also be a valuable resource in the context of predictive models. In this section we present several case studies to demonstrate how the proposed perturbations can be useful for systematically evaluating predictive robustness under controlled domain shift and how morphometry enables an entirely new variety of experimental settings and nuanced analyses based on shape stratification. Further we demonstrate how the perturbation labels and morphometrics can serve as prediction targets in their own right for new supervised tasks. For simplicity and reproducibility, we use the same datasets as in the disentanglement experiments (Section 6.4.2): plain digits (PLAIN), plain mixed with thinned and thickened digits (GLOBAL), and plain mixed with swollen and fractured digits (LOCAL).

6.4.3.1 Robustness

Synthetic perturbations enable fabrication of domain shift scenarios, aiming to systematically evaluate the robustness of predictive models to specific types of variation. For example, digits with local perturbations (swelling and fractures) are still clearly identifiable by humans (see Figs. B.8 and B.9). However, a practitioner might be interested in investigating how different types of classifiers react to these changes.

Here we compare simple representative methods from four classification paradigms: neighbourhood-based, kernel-based, fully connected, and convolutional. Specifically, we evaluated k-nearest-neighbours (kNN) using $k = 5$ neighbours and ℓ_1 distance weighting, a support vector machine (SVM) with polynomial kernel and penalty parameter $C = 100$, a multi-layer perceptron (MLP) with 784–200–200–L architecture (L: number of outputs), and a LeNet-5 (LeCun et al., 1998) convolutional neural network (CNN).

Each model is trained for digit recognition once on PLAIN, then tested separately on both PLAIN and LOCAL test datasets, to investigate the effect of *domain shift*. All methods suffer a drop in test accuracy on LOCAL (Table 6.3), and, considering that one third of the images in LOCAL are identical to ones in PLAIN, the absolute performance drops for the perturbed digits alone are in fact in the order of 1.5%, 4.9%, 7.2%, and 5.4%, respectively. kNN appears to be the most robust to local perturbations, conceivably because they affect only a few pixels, leaving the image distance between neighbours largely unchanged. Contrarily, both the global patterns that SVM and MLP rely on and the CNN's local filter responses may have changed considerably. Such analyses can provide new insights into the behaviour, strengths, and weaknesses of different models, and can be useful to inform new strategies for developing more robust approaches.

Test data	*k*NN	SVM	MLP	CNN
PLAIN	96.25	95.71	97.97	98.95
LOCAL	95.22	92.47	93.15	95.33

Table 6.3 Example results on digit recognition (accuracy, %) under synthetic domain shift using the proposed data perturbations

6.4.3.2 Data stratification

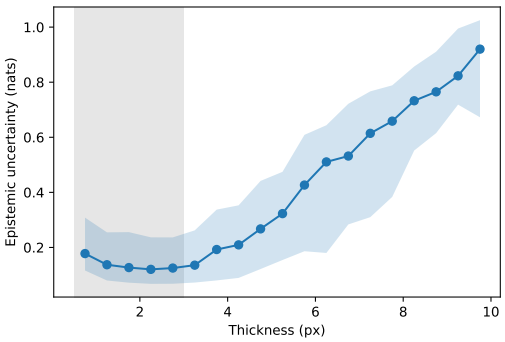
Originally, MNIST could be stratified exclusively by digit class, as this was the only meta-information available. However, with morphometrics, data can be sliced and grouped according to entirely new criteria: we are able to select the shape properties we desire in the training set and also perform finer analyses at test time. It is even possible to combine multiple attributes for multifaceted studies.

To illustrate how these capabilities can be put to use, we present a simplified case study on a hypothetical unsupervised outlier detection scenario. We take the GLOBAL dataset, restrict the training data to digits whose thickness is below 3 px, and evaluate the model's response to test digits across the entire thickness range, especially out-of-distribution samples. Here we assume the true selection criterion is unknown at learning time. Two example frameworks were considered: uncertainty-based and reconstruction-based.

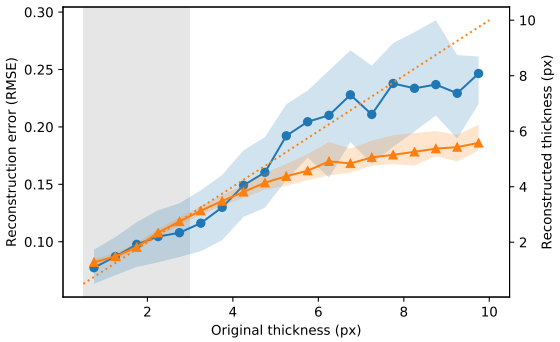
For the first setting, we assume a research question about outlier detection with respect to the digit classification task. We take a LeNet-5 classifier (LeCun et al., 1998) and employ Monte Carlo dropout (Kendall and Gal, 2017) to estimate the model's epistemic uncertainty via mutual information (L. Smith and Gal, 2018). This metric, which quantifies the model's uncertainty about its parameters when presented with a test input, has been shown to be effective at detecting out-of-distribution samples (L. Smith and Gal, 2018). Figure 6.10a shows that the model seems markedly more uncertain on average regarding thicker digits (e.g. above 5 px), although there is no abrupt jump in uncertainty at the inlier boundary. Since we have 'ground-truth' outlier labels for the test images based on their thickness, we are also able to evaluate the discriminative power of this score: with an area under ROC curve (AUC) of only 0.67, this approach did not result in a particularly strong outlier detector.

In the second setting, we explored outlier detection with a deep convolutional auto-encoder.[38] It was likewise trained on digits thinner than 3 px, and we plot in Fig. 6.10b its reconstruction error for test digits of varying thicknesses. We observe a steeper trend than for the previous model's uncertainty (Fig. 6.10a), and in fact this auto-encoder's reconstruction error is far more discriminative (AUC = 0.85). In addition, it is also possible to collect morphometrics of the reconstructions and visualise to what extent the original shape attributes are preserved. Figure 6.10b suggests the interesting finding that the reconstructions of thick digits are consistently thinner than the originals, as if the auto-encoder

[38] The encoder has three convolutional and one fully connected layers, and the mirrored decoder employs transposed convolutions. The convolutional layers have 3×3 filters with stride 2 and 32–64–128 feature channels, and the bottleneck layer is 10 units wide.



a Uncertainty of MC-dropout model



b Auto-encoder reconstruction error (left, ●) and thickness of reconstructed digits (right, ▲)

Figure 6.10 Results for toy outlier detection experiments showcasing data stratification by stroke thickness. Training digits were restricted to those thinner than 3 px (shaded range), and test digits were grouped into 0.5 px-wide bins for analysis. Lines and shaded regions indicate medians and interquartile ranges.

Table 6.4 Example results on some of the new MNIST prediction tasks

Prediction task	*k*NN	SVM	MLP	CNN
Abnormality detection (%)	65.10	77.59	88.25	97.53
Thickness regression (RMSE)	0.467	0.365	0.348	0.279

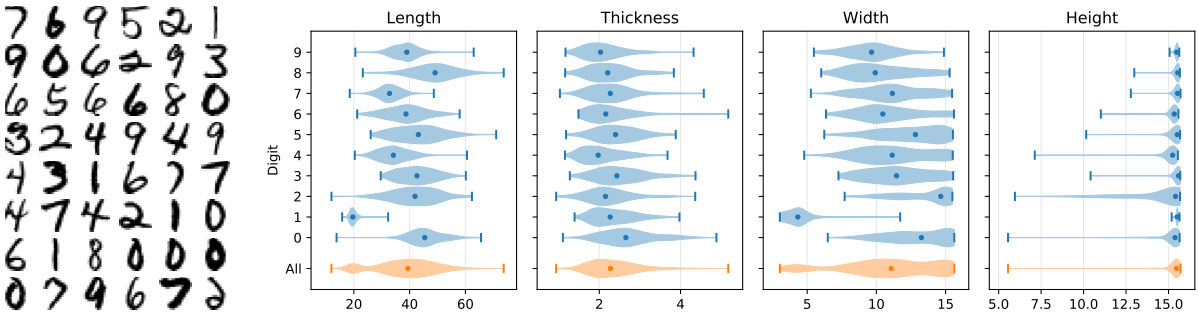
'projected' these digits towards the training subspace to some extent.

6.4.3.3 New prediction tasks

In the previous two sections, we demonstrated how perturbations may be leveraged to systematically evaluate robustness to domain shift and how morphometry can prove valuable for selecting training data and for performing stratified test-time analysis. Here we illustrate how perturbation labels and morphometric attributes can be used for interesting new prediction tasks on MNIST, evaluating the same models from Section 6.4.3.1.

As our first example, we define a supervised abnormality detection task, using the LOCAL dataset, to predict whether a digit is normal or perturbed (swollen or fractured)—compare with lesion detection in medical scans. The top row of Table 6.4 indicates that the CNN was able to detect abnormalities with high accuracy, likely thanks to local invariances of its architecture, whereas the shallow classifiers, *k*NN and SVM, performed remarkably worse. Note how the performance is generally much lower than for digit classification, revealing the higher difficulty of this binary problem compared to the ten-class recognition task.

Finally, we also constructed a regression task for digit thickness using the GLOBAL dataset, mimicking medical imaging tasks such as estimating brain age from cortical grey matter maps. Since this is a non-trivial task requiring some awareness of local geometry, it is unsurprising that the convolutional



a USPS digits **b** Distribution of morphological attributes for USPS test images

Figure 6.11 Application of the proposed morphometrics to the USPS dataset

model outperformed the others, which rely on holistic features (Table 6.4, bottom row).

6.5 Discussion and perspectives

Through concrete examples, we demonstrated in Section 6.4 how morphometry and perturbations can be exploited 1) for qualitative and, more importantly, *quantitative* evaluation of generative models in terms of sample diversity and disentanglement, and 2) in developing novel experimental protocols for analysing predictive models. The presented methodologies are broadly applicable to datasets of rasterised shapes beyond MNIST.

For example, we apply our exact same morphometry pipeline to handwritten digits from the USPS dataset (LeCun et al., 1990), also composed of small grayscale images (16×16). Example images from this dataset and distributions of length, thickness, width, and height can be seen in Fig. 6.11. Slant was excluded because the USPS digits were already deslanted, with residual variation in the order of $\pm 1^\circ$. Another popular dataset to which our morphometrics can be directly applied is Omniglot (Lake et al., 2015), consisting of handwritten symbols from 50 different alphabets.

Besides, the general perspective of measuring attributes directly from images is evidently not limited to datasets of handwritten characters. Consider for example the dSprites dataset (Matthey et al., 2017), often used in disentanglement research, which contains 64×64 synthetic binary images of small shapes with ground-truth annotations. Attributes such as location and scale can easily be measured based on image moments (see Section 6.2.4). They are expected to be robust even for blurry generated samples. As discussed previously, while the annotations provided with such data facilitate evaluation of inferential disentanglement, these measurements enable the unprecedented analysis of generative behaviour.

An obvious limitation of morphometry is the tacit assumption that generated samples resemble true data closely enough that the shape measurements are indeed meaningful. What happens if the

generative model under study is underfitted and produces nonsensical images? In such case, one expects the joint distribution of sample morphometrics to be strikingly different from that of true data, even if some individual metrics may have been extracted successfully (e.g. width and height of a reasonably sized blur). Like mode collapse, this issue can be diagnosed straightforwardly—for example with an analysis similar to that from Section 6.4.1—and may constitute an interesting finding by itself.

6.6 Conclusion

With Morpho-MNIST we provide a number of mechanisms to quantitatively assess representation learning with respect to measurable factors of variation in the data. We believe that this may prove an important asset for future research on generative models, and will hopefully also be useful in evaluating predictive models. We emphasise that the proposed morphometrics can be used *post hoc* to evaluate already trained models, potentially revealing novel insights about their behaviour. In addition, we have shown that our morphometry and perturbation framework easily extends to datasets other than MNIST. We hope that the case studies presented here may also invite research into developing metrics for other kinds of images beyond rasterised 2D shapes or even to non-imaging data, paving the way to more widespread analysis of true generative performance.

Having developed this fully controllable experimental environment with images and non-imaging covariates, we return to the task of building probabilistic models that account for causal structure. The next chapter dives into structural causal models, a complete functional specification of a data-generating process that can answer queries at all levels of the causal hierarchy (Section 2.3.1). In particular, we propose to implement the causal mechanisms generating high-dimensional data by using deep-learning components. The choices we present enable tractable prediction of counterfactuals in accordance with causality theory, which to our knowledge has not previously been achieved in the literature.

Chapter 7

Structural causal models for images and non-imaging covariates

This chapter is based on the following publication:

- Nick Pawlowski*, **Daniel C. Castro*** and Ben Glocker (2020). 'Deep structural causal models for tractable counterfactual inference'. In: *Advances in Neural Information Processing Systems (NeurIPS 2020)*. Forthcoming, *equal contribution. arXiv: 2006.06485.

This is a shared first-authorship with equal contribution from Nick Pawlowski and Daniel C. Castro. Both authors contributed to all aspects of the work with main responsibilities as follows: N.P. led the implementation of specific models, data generation and processing, and experimentation; D.C.C. had main responsibilities on the framework design, analysis, and background research. Both took equal responsibility on ideation, conceptualisation, framework implementation, and writing of the manuscript.
The code for all experiments and a link to an interactive demonstration are available at https://github.com/biomedia-mira/deepscm.

In this final part of the thesis, we formulate a general framework for building structural causal models (SCMs) with deep learning components. The proposed approach employs normalising flows and variational inference to enable tractable inference of exogenous noise variables—a crucial step for counterfactual inference that is missing from existing deep causal learning methods. Our framework is validated on a synthetic dataset based on Morpho-MNIST (Chapter 6) as well as on a real-world medical dataset of brain MRI scans. Our experimental results indicate that we can successfully train deep SCMs that are capable of all three levels of Pearl's ladder of causation: association, intervention, and counterfactuals, giving rise to a powerful new approach for answering causal questions in imaging applications and beyond.

7.1 Introduction

Many questions in everyday life as well as in scientific inquiry are causal in nature: "How would the climate have changed if we'd had less emissions in the '80s?", "How fast could I run if I hadn't been smoking?", or "Will my headache be gone if I take that pill?". None of those questions can be answered with statistical tools alone, but require methods from causality to analyse interactions with our environment (interventions) and hypothetical alternate worlds (counterfactuals), going beyond joint, marginal, and conditional probabilities (J. Peters et al., 2017). Even though these are natural lines of reasoning, their mathematical formalisation under a unified theory is relatively recent (Pearl, 2009).

In some statistics-based research fields, such as econometrics or epidemiology, the use of causal inference methods has been established for some time (Greenland et al., 1999; Wold, 1954). However, causal approaches have been introduced into deep learning (DL) only very recently (Schölkopf, 2019). For example, research has studied the use of causality for disentanglement (Parascandolo et al., 2018; Yang et al., 2020), causal discovery (Bengio et al., 2020; Goudet et al., 2018), and for deriving causality-inspired explanations (Martínez and Marca, 2019; Singla et al., 2020) or data augmentations (Ilse et al., 2020; Kaushik et al., 2020). Causal DL models could be capable of learning relationships from complex high-dimensional data and of providing answers to interventional and counterfactual questions, although current work on deep counterfactuals is limited by modelling only direct cause-effect relationships (Singla et al., 2020) or instrumental-variable scenarios (Hartford et al., 2017), or by not providing a full recipe for tractable counterfactual inference (Kocaoglu et al., 2018).

The integration of causality into DL research promises to enable novel scientific advances as well as to tackle known shortcomings of DL methods: DL is known to be susceptible to learning spurious correlations and amplifying biases (e.g. J. Zhao et al., 2017), and to be exceptionally vulnerable to changes in the input distribution (Szegedy et al., 2014). By explicitly modelling causal relationships and acknowledging the difference between causation and correlation, causality becomes a natural field of study for improving the transparency, fairness, and robustness of DL-based systems (Kusner et al., 2017; Subbaswamy et al., 2019). Further, the tractable inference of deep counterfactuals enables novel research avenues that aim to study causal reasoning on a per-instance rather than population level, which could lead to advances in personalised medicine as well as in decision-support systems, more generally.

In this context, our work studies the use of DL-based causal mechanisms and establishes effective ways of performing counterfactual inference. Our main contributions are: 1) a unified framework for structural causal models using modular deep mechanisms; 2) an efficient approach to estimating counterfactuals by inferring exogenous noise via variational inference or normalising flows; 3) case

studies exemplifying how to apply deep structural causal models and perform counterfactual inference. The chapter is organised as follows: we first review structural causal models and discuss how to leverage deep mechanisms and enable tractable counterfactual inference. Second, we compare our work to recent progress in deep causal learning in light of Pearl's ladder of causation (Pearl, 2019). Finally, we apply deep structural causal models to a synthetic experiment as well as to modelling brain MRI scans, demonstrating the practical utility of our framework in answering counterfactual questions.

7.2 Deep structural causal models

We consider the problem of modelling a collection of K random variables $\mathbf{x} = (x_1, \ldots, x_K)$. By considering causal relationships between them, we aim to build a model that not only is capable of generating convincing novel samples, but also satisfies all three rungs of the causal hierarchy (Section 2.3.1). The three levels of causation can be operationalised by employing structural causal models (SCMs)[39], recapitulated in the next section.

7.2.1 Structural causal models

A structural causal model $\mathfrak{G} := (\mathbf{S}, P(\boldsymbol{\epsilon}))$ consists of a collection $\mathbf{S} = (f_1, \ldots, f_K)$ of structural assignments $x_k := f_k(\epsilon_k; \mathbf{pa}_k)$ (called *mechanisms*), where $\mathbf{pa}_k$ is the set of parents of x_k (its *direct causes*), and a joint distribution $P(\boldsymbol{\epsilon}) = \prod_{k=1}^{K} P(\epsilon_k)$ over mutually independent exogenous noise variables (i.e. unaccounted sources of variation). As assignments are assumed acyclic, relationships can be represented by a directed acyclic graph (DAG) with edges pointing from causes to effects, called the *causal graph* induced by $\mathfrak{G}$. Every SCM $\mathfrak{G}$ entails a unique joint observational distribution $P_{\mathfrak{G}}(\mathbf{x})$, satisfying the causal Markov assumption: each variable is independent of its non-effects given its direct causes. It therefore factorises as $P_{\mathfrak{G}}(\mathbf{x}) = \prod_{k=1}^{K} P_{\mathfrak{G}}(x_k \mid \mathbf{pa}_k)$, where each conditional distribution $P_{\mathfrak{G}}(x_k \mid \mathbf{pa}_k)$ is determined by the corresponding mechanism and noise distribution (J. Peters et al., 2017).

Crucially, unlike conventional Bayesian networks, the conditional factors above are imbued with a causal interpretation. This enables $\mathfrak{G}$ to be used to predict the effects of *interventions*, defined as substituting one or multiple of its structural assignments, written as 'do(…)'. In particular, a constant reassignment of the form $\text{do}(x_k := a)$ is called an atomic intervention, which disconnects x_k from all its parents and represents a direct manipulation disregarding its natural causes.

While the observational distribution relates to statistical associations and interventions can predict causal effects, SCMs further enable reasoning about *counterfactuals*. These are hypothetical

[39] SCMs are also known as (nonlinear) structural equation models or functional causal models.

retrospective interventions, given an observed outcome: 'What would x_i have been if x_j were different, given that we observed $\mathbf{x}$?'. This type of question effectively offers explanations of the data, since we can analyse the changes resulting from manipulating each variable. Counterfactual queries can be mathematically formulated as a three-step procedure (Pearl, 2009, Ch. 7):

1. **Abduction:** Predict the 'state of the world' (the exogenous noise, $\boldsymbol{\epsilon}$) that is compatible with the observations, $\mathbf{x}$, i.e. infer $P_{\mathfrak{G}}(\boldsymbol{\epsilon} \mid \mathbf{x})$.

2. **Action:** Perform an intervention (e.g. $\text{do}(x_k := \widetilde{x}_k)$) corresponding to the desired manipulation, resulting in a modified SCM $\widetilde{\mathfrak{G}} = \mathfrak{G}_{\mathbf{x};\text{do}(\widetilde{x}_k)} = (\widetilde{\mathbf{S}}, P_{\mathfrak{G}}(\boldsymbol{\epsilon} \mid \mathbf{x}))$ (J. Peters et al., 2017, Sec. 6.4).

3. **Prediction:** Compute the quantity of interest based on the distribution entailed by the counterfactual SCM, $P_{\widetilde{\mathfrak{G}}}(\mathbf{x})$.

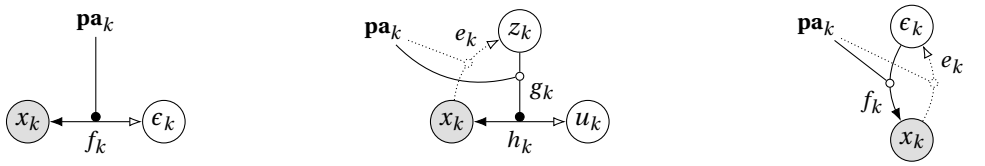
With these operations in mind, the next section explores a few options for building flexible, expressive, and counterfactual-capable functional mechanisms for highly structured data.

7.2.2 Deep mechanisms

In statistical literature (e.g. epidemiology, econometrics, sociology), SCMs are typically employed with simple linear mechanisms (or generalised linear models, involving an output non-linearity). Analysts attach great importance to the regression weights, as under certain conditions these may be readily interpreted as estimates of the causal effects between variables. While this approach generally works well for scalar variables and can be useful for decision-making, it is not flexible enough to model higher-dimensional data such as images. Solutions to this limitation have been proposed by introducing deep-learning techniques into causal inference (Goudet et al., 2018; Kocaoglu et al., 2018).

We call an SCM that uses deep-learning components to model the structural assignments a *deep structural causal model* (DSCM). In DSCMs, the inference of counterfactual queries becomes more complex due to the potentially intractable abduction step (inferring the posterior noise distribution, as defined above). To overcome this, we propose to use recent advances in normalising flows and variational inference to model mechanisms for composable DSCMs that enable tractable counterfactual inference. While here we focus on continuous data, DSCMs also fully support discrete variables without the need for relaxations (see Section 7.2.4). We consider three types of mechanisms that differ mainly in their invertibility, illustrated in Fig. 7.1.

Invertible, explicit: Recall from Section 2.2.3 that normalising flows model complex probability distributions using invertible transformations from simpler base distributions. This class of models



a Invertible explicit likelihood **b** Amortised explicit likelihood **c** Amortised implicit likelihood

Figure 7.1 Classes of deep causal mechanisms considered in this work. Bi-directional arrows indicate invertible transformations, optionally conditioned on other inputs (edges ending in black circles). Black and white arrowheads refer resp. to the generative and abductive directions, while dotted arrows depict an amortised variational approximation. Here, f_k is the forward model, e_k is an encoder that amortises abduction in non-invertible mechanisms, g_k is a 'high-level' non-invertible branch (e.g. a probabilistic decoder), and h_k is a 'low-level' invertible mapping (e.g. reparametrisation).

also subsumes the typical location-scale and inverse cumulative distribution function transformations used in the reparametrisation trick (Kingma and Welling, 2014a; Rezende et al., 2014), as well as the Gumbel trick for discrete variable relaxations (Jang et al., 2017; Maddison et al., 2017).

Although normalising flows were originally proposed for unconditional distributions, they have been extended to conditional densities (Trippe and R. E. Turner, 2017), including in high dimensions (Lu and B. Huang, 2020; Winkler et al., 2019), by parametrising the transformation as $x = f(\epsilon; \mathbf{pa}_X)$, assumed invertible in the first argument. In particular, conditional flows can be adopted in DSCMs to represent invertible, explicit-likelihood mechanisms (Fig. 7.1a):

$$x_i \coloneqq f_i(\epsilon_i; \mathbf{pa}_i), \qquad p(x_i \,|\, \mathbf{pa}_i) = p(\epsilon_i) \cdot \left|\det \nabla_{\epsilon_i} f_i(\epsilon_i; \mathbf{pa}_i)\right|^{-1} \Big|_{\epsilon_i = f_i^{-1}(x_i; \mathbf{pa}_i)}. \tag{7.1}$$

Amortised, explicit: Such invertible architectures typically come with heavy computational requirements when modelling high-dimensional observations, because all intermediate operations act in the same dimensionality as the data. Instead, it is possible to use arbitrary functional forms for the structural assignments, at the cost of losing invertibility and tractable likelihoods $p(x_k \,|\, \mathbf{pa}_k)$. Here, we propose to separate the assignment f_k into a 'low-level', invertible component h_k and a 'high-level', non-invertible part g_k—with a corresponding noise decomposition $\epsilon_k = (u_k, z_k)$—such that

$$x_k \coloneqq f_k(\epsilon_k; \mathbf{pa}_k) = h_k(u_k; g_k(z_k; \mathbf{pa}_k), \mathbf{pa}_k), \qquad P(\epsilon_k) = P(u_k)P(z_k)\,. \tag{7.2}$$

In such a decomposition, the invertible transformation h_k can be made shallower, while the upstream non-invertible g_k maps from a lower-dimensional space and is expected to capture more of the high-level structure of the data. Indeed, a common implementation of this type of model for images would involve a probabilistic decoder, where g_k may be a convolutional neural network, predicting the parameters of a simple location-scale transformation performed by h_k (Kingma and Welling, 2014a).

As the conditional likelihood $p(x_k \,|\, \mathbf{pa}_k)$ in this class of models is no longer tractable because

z_k cannot be marginalised out, it may alternatively be trained with amortised variational inference (Section 2.2.2.2). Specifically, we can introduce a variational distribution $Q(z_k \mid x_k, \mathbf{pa}_k)$ to formulate a lower bound on the true marginal conditional log-likelihood, which will be maximised instead:

$$\log p(x_k \mid \mathbf{pa}_k) \geq \mathbb{E}_{Q(z_k \mid x_k, \mathbf{pa}_k)}[\log p(x_k \mid z_k, \mathbf{pa}_k)] - D_{\mathrm{KL}}[Q(z_k \mid x_k, \mathbf{pa}_k) \| P(z_k)]\,. \tag{7.3}$$

The argument of the expectation in this lower bound can be calculated similarly to Eq. (7.1):

$$p(x_k \mid z_k, \mathbf{pa}_k) = p(u_k) \cdot \left|\det \nabla_{u_k} h_k(u_k; g_k(z_k, \mathbf{pa}_k), \mathbf{pa}_k)\right|^{-1} \Big|_{u_k = h_k^{-1}(x_k; g_k(z_k, \mathbf{pa}_k), \mathbf{pa}_k)}\,. \tag{7.4}$$

The approximate posterior distribution $Q(z_k \mid x_k, \mathbf{pa}_k)$ can for example be realised by an encoder function, $e_k(x_k; \mathbf{pa}_k)$, that outputs the parameters of a simple distribution over z_k (Fig. 7.1b).

Amortised, implicit: While the models above rely on (approximate) maximum-likelihood as training objective, it is admissible to train a non-invertible mechanism as a conditional implicit-likelihood model (Fig. 7.1c), optimising an adversarial objective (Donahue et al., 2017; Dumoulin et al., 2017; Mirza and Osindero, 2014). Specifically, a deterministic encoder e_j would strive to fool a discriminator function attempting to tell apart tuples of encoded real data $(x_j, e_j(x_j; \mathbf{pa}_j), \mathbf{pa}_j)$ and generated samples $(f_j(\epsilon_j; \mathbf{pa}_j), \epsilon_j, \mathbf{pa}_j)$.

Though we did not directly experiment with this approach in Pawlowski*, **Castro*** et al. (2020), it has since been attempted by Dash and Sharma (2020). Leveraging such an adversarially trained image mechanism, they were able to achieve comparable results in counterfactual image generation to the conditional-VAE-based one we evaluated on Morpho-MNIST (Section 7.4).

7.2.3 Deep counterfactual inference

Now equipped with effective deep models for representing mechanisms in DSCMs, we discuss the inference procedure allowing us to compute answers to counterfactual questions.

Abduction: As presented in Section 7.2.1, the first step in computing counterfactuals is abduction, i.e. to predict the exogenous noise, $\boldsymbol{\epsilon}$, based on the available evidence, $\mathbf{x}$. Because each noise variable is assumed to affect only the respective observed variable, $(\epsilon_k)_{k=1}^K$ are conditionally independent given $\mathbf{x}$, therefore this posterior distribution factorises as

$$P_{\mathfrak{G}}(\boldsymbol{\epsilon} \mid \mathbf{x}) = \prod_{k=1}^{K} P_{\mathfrak{G}}(\epsilon_k \mid x_k, \mathbf{pa}_k)\,. \tag{7.5}$$

In other words, it suffices to infer the noise independently for each mechanism, given the observed values of the variable and of its parents.[40]

For invertible mechanisms, the noise variable can be obtained deterministically and exactly by just inverting the mechanism: $\epsilon_i = f_i^{-1}(x_i; \mathbf{pa}_i)$. Similarly, implicit-likelihood mechanisms can be approximately inverted by using the trained encoder function: $\epsilon_j \approx e_j(x_j; \mathbf{pa}_j)$.

Some care must be taken in the case of amortised, explicit-likelihood mechanisms, as the 'high-level' noise z_k and 'low-level' noise u_k are not independent given x_k. Recalling that this mechanism is trained along with a conditional probabilistic encoder, $Q(z_k \mid e_k(x_k; \mathbf{pa}_k))$, the noise posterior can be approximated as follows, where $\delta_w(\cdot)$ denotes the Dirac delta distribution centred at w:

$$\begin{aligned} P_{\mathfrak{G}}(\epsilon_k \mid x_k, \mathbf{pa}_k) &= P_{\mathfrak{G}}(z_k \mid x_k, \mathbf{pa}_k)\, P_{\mathfrak{G}}(u_k \mid z_k, x_k, \mathbf{pa}_k) \\ &\approx Q(z_k \mid e_k(x_k; \mathbf{pa}_k))\, \delta_{h_k^{-1}(x_k; g_k(z_k; \mathbf{pa}_k), \mathbf{pa}_k)}(u_k)\,. \end{aligned} \tag{7.6}$$

Action: The causal graph is then modified according to the desired hypothetical intervention(s), as in the general case (Section 7.2.1). For each intervened variable x_k, its structural assignment is replaced either by a constant, $x_k := \tilde{x}_k$—making it independent of its former parents (direct causes, $\mathbf{pa}_k$) and of its exogenous noise (ϵ_k)—or by a surrogate mechanism $x_k := \widetilde{f}_k(\epsilon_k; \widetilde{\mathbf{pa}}_k)$, forming a set of counterfactual assignments, $\widetilde{\mathbf{S}}$. This then defines a counterfactual SCM $\widetilde{\mathfrak{G}} = (\widetilde{\mathbf{S}}, P_{\mathfrak{G}}(\boldsymbol{\epsilon} \mid \mathbf{x}))$.

Prediction: Finally, we can sample from $\widetilde{\mathfrak{G}}$. Noise variables that were deterministically inverted (either exactly or approximately) can simply be plugged back into the respective forward mechanism to determine the new output value. Notice that this step is redundant for observed variables that are not descendants of the ones being intervened upon, as they will be unaffected by the changes.

As mentioned above, the posterior distribution over (z_k, u_k) for an amortised, explicit-likelihood mechanism does not factorise (Eq. (7.6)), and the resulting distribution over the counterfactual x_k cannot be characterised explicitly. However, sampling from it is straightforward, such that we can approximate the counterfactual distribution via Monte Carlo as follows, for each sample s:

$$z_k^{(s)} \sim Q(z_k \mid e_k(x_k; \mathbf{pa}_k)) \tag{7.7a}$$

$$u_k^{(s)} = h_k^{-1}(x_k; g_k(z_k^{(s)}; \mathbf{pa}_k), \mathbf{pa}_k) \tag{7.7b}$$

$$\tilde{x}_k^{(s)} = \widetilde{h}_k(u_k^{(s)}; \widetilde{g}_k(z_k^{(s)}; \widetilde{\mathbf{pa}}_k), \widetilde{\mathbf{pa}}_k)\,. \tag{7.7c}$$

[40] Note that here we assume full observability, i.e. no variables are missing when predicting counterfactuals. We discuss challenges of handling partial evidence in Section 7.6 and Section 8.2.2.

Consider an uncorrelated Gaussian decoder for images as a concrete example, predicting vectors of means and variances for each pixel of x_k: $g_k(z_k;\mathbf{pa}_k) = (\mu(z_k;\mathbf{pa}_k), \sigma^2(z_k;\mathbf{pa}_k))$. Exploiting the reparametrisation trick, counterfactuals that preserve x_k's mechanism can be computed simply as

$$u_k^{(s)} = [x_k - \mu(z_k^{(s)};\mathbf{pa}_k)] \oslash \sigma(z_k^{(s)};\mathbf{pa}_k) \tag{7.8a}$$

$$\widetilde{x}_k^{(s)} = \mu(z_k^{(s)};\widetilde{\mathbf{pa}}_k) + \sigma(z_k^{(s)};\widetilde{\mathbf{pa}}_k) \odot u_k^{(s)}\,, \tag{7.8b}$$

where $\oslash$ and $\odot$ denote element-wise division and multiplication, respectively. In particular, in the constant-variance setting adopted for our experiments, counterfactuals further simplify to

$$\widetilde{x}_k^{(s)} = \mu(z_k^{(s)};\widetilde{\mathbf{pa}}_k) + \underbrace{[x_k - \mu(z_k^{(s)};\mathbf{pa}_k)]}_{\substack{\text{residuals of original image,}\\ \text{added to counterfactual mean}}} = x_k + \underbrace{[\mu(z_k^{(s)};\widetilde{\mathbf{pa}}_k) - \mu(z_k^{(s)};\mathbf{pa}_k)]}_{\substack{\text{counterfactual 'correction' term,}\\ \text{added to original image}}}. \tag{7.9}$$

This showcases how true image counterfactuals are able to retain pixel-level details. Typical conditional generative models would output only $\mu(z_k;\widetilde{\mathbf{pa}}_k)$ (which is often blurry in vanilla variational auto-encoders (Larsen et al., 2016)), or would in addition have to sample $P(u_k)$ (resulting in noisy images).

7.2.4 Discrete counterfactuals

As mentioned earlier, the DSCM framework supports not only low- and high-dimensional continuous data, but also discrete variables. In particular, discrete mechanisms with a Gumbel–max parametrisation were shown to lead to counterfactuals satisfying desirable properties (Oberst and Sontag, 2019). For example, they are invariant to category permutations and are stable, such that increasing the odds only of the observed outcome cannot produce a different counterfactual outcome. More computational details and properties of the Gumbel distribution are found in Maddison and Tarlow (2017).

Consider a discrete random variable over K categories, y, with a conditional likelihood described by logits $\boldsymbol{\lambda}$, assumed to be a function g_Y of its parents, $\mathbf{pa}_Y$:

$$P(y = k\,|\,\mathbf{pa}_Y) = \frac{e^{\lambda_k}}{\sum_{l=1}^{K} e^{\lambda_l}}, \qquad \boldsymbol{\lambda} = g_Y(\mathbf{pa}_Y)\,. \tag{7.10}$$

Under the Gumbel–max parametrisation, the mechanism generating y can be described as

$$y := f_Y(\boldsymbol{\epsilon}_Y;\mathbf{pa}_Y) = \underset{1\le l\le K}{\arg\max}\,(\epsilon_Y^l + \lambda_l), \qquad \epsilon_Y^l \sim \text{Gumbel}(0,1)\,. \tag{7.11}$$

Samples from the Gumbel$(0,1)$ distribution can be generated by computing $-\log(-\log U)$, where $U \sim \text{Unif}(0,1)$.

The Gumbel distribution has certain special properties (Maddison and Tarlow, 2017) that enable tractable abduction. Given that we observed $y = k$, samples can be generated from the exact posterior $P(\boldsymbol{\epsilon}_Y \mid y = k, \mathbf{pa}_Y)$:

$$\begin{aligned} \epsilon_Y^k &= G_k + \log \textstyle\sum_l e^{\lambda_l} - \lambda_k, & G_k &\sim \text{Gumbel}(0,1), \\ \epsilon_Y^l &= -\log(e^{-G_l - \lambda_l} + e^{-\epsilon_Y^k - \lambda_k}) - \lambda_l, & G_l &\sim \text{Gumbel}(0,1), \quad \forall l \neq k. \end{aligned} \tag{7.12}$$

Finally, given an upstream counterfactual intervention such that $\widetilde{\boldsymbol{\lambda}} = \widetilde{g}_Y(\widetilde{\mathbf{pa}}_Y)$, the counterfactual outcome for y can be determined simply as

$$\widetilde{y} = f_Y(\boldsymbol{\epsilon}_Y; \widetilde{\mathbf{pa}}_Y) = \underset{1 \leq l \leq K}{\arg\max} (\epsilon_Y^l + \widetilde{\lambda}_l). \tag{7.13}$$

Note that this entire derivation applies to a truly discrete variable, without the need for continuous relaxations as commonly used in deep generative models (Jang et al., 2017; Maddison et al., 2017), as the likelihood is given in closed form and no gradients of expectations are necessary.

7.3 Related work

Deep generative modelling has seen a wide range of contributions since the popularisation of variational auto-encoders (VAEs; Kingma and Welling, 2014a), generative adversarial networks (GANs; Goodfellow et al., 2014), and normalising flows (Rezende and Mohamed, 2015). These models have since been employed to capture conditional distributions (Mirza and Osindero, 2014; Sohn et al., 2015; Trippe and R. E. Turner, 2017; Winkler et al., 2019), and VAEs and GANs were also extended to model structured data by incorporating probabilistic graphical models (Johnson et al., 2016; C. Li et al., 2018; Lin et al., 2018). In addition, deep generative models have been heavily used for (unsupervised) representation learning with an emphasis on disentanglement (X. Chen et al., 2016; Higgins et al., 2017; Kulkarni et al., 2015; Pati and Lerch, 2020). However, even when these methods faithfully capture the distribution of observed data, they are capable of fulfilling only the association rung of the ladder of causation.

Interventions build on the associative capabilities of probabilistic models to enable queries related to changes in causal mechanisms. By integrating a causal graph into the connectivity of a deep model, it is possible to perform interventions with GANs (Kocaoglu et al., 2018) and causal generative NNs (Goudet et al., 2018). VAEs can also express causal links using specific covariance matrices between

latent variables, which however restrict the dependences to be linear (Yang et al., 2020). Despite reaching the second rung of the causal ladder, these methods lack tractable abduction capabilities and therefore cannot generate counterfactuals.

Some machine-learning tasks such as explainability, image-to-image translation, or style transfer are closely related to counterfactual queries of the sort 'How would x (have to) change if we (wished to) modify y?'. Here, y could be the style of a picture for style transfer (Gatys et al., 2016), the image domain (e.g. drawing to photo) for image-to-image translation (Isola et al., 2017), the age of a person in natural images (Antipov et al., 2017) or medical scans (Xia et al., 2019), or a predicted output for explainability (Singla et al., 2020). However, these approaches do not explicitly model associations, interventions, nor causal structure. Potentially closest to our work is a method for counterfactual explainability of visual models, which extends CausalGANs (Kocaoglu et al., 2018) to predict reparametrised distributions over image attributes following an assumed causal graph (Martínez and Marca, 2019). However, this approach performs no abduction step, instead resampling the noise of attributes downstream from the intervention(s), and does not include a generative model of imaging data. To the best of our knowledge, the proposed DSCM framework is the first flexible approach enabling end-to-end training and tractable inference on all three levels of the ladder of causation for high-dimensional data.

7.4 Case study 1: Morpho-MNIST

We consider the problem of modelling the causal model of a synthetic dataset based on MNIST digits (LeCun et al., 1998), where stroke thickness causes the brightness of the digit: thicker digits are brighter whereas thinner digits are dimmer. This simple dataset allows for examining the three levels of causation in a controlled and measurable environment. We use morphological transformations on MNIST (Chapter 6) to generate a dataset with known causal structure and access to the 'true' process of generating counterfactuals.

7.4.1 Data generation

We define the SCM for generating this synthetic dataset as follows:

$$t := f_T^*(\epsilon_T^*) = 0.5 + \epsilon_T^*, \qquad \epsilon_T^* \sim \Gamma(10,5)\,. \tag{7.14a}$$

$$i := f_I^*(\epsilon_I^*; t) = 191 \cdot \sigma(0.5 \cdot \epsilon_I^* + 2 \cdot t - 5) + 64, \qquad \epsilon_I^* \sim \mathcal{N}(0,1)\,. \tag{7.14b}$$

$$x := f_X^*(\epsilon_X^*; i, t) = \mathsf{SetIntensity}(\mathsf{SetThickness}(\epsilon_X^*; t); i), \qquad \epsilon_X^* \sim \mathrm{MNIST}\,, \tag{7.14c}$$

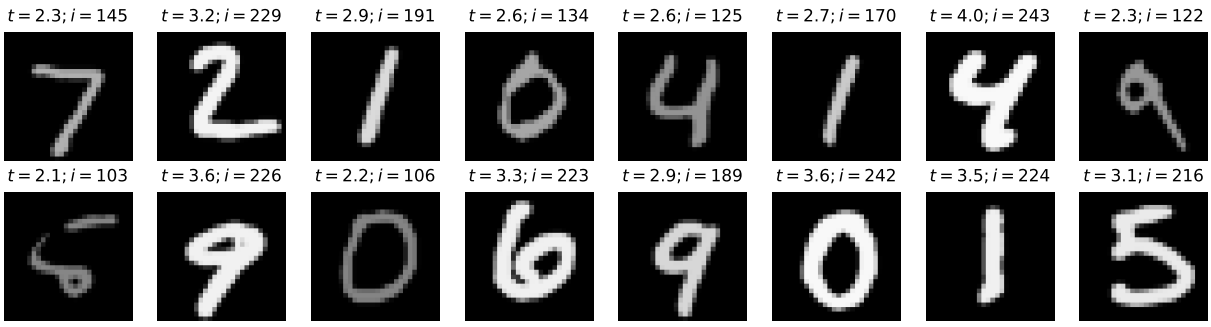


Figure 7.2 Random exemplars from the synthetically generated Morpho-MNIST test dataset

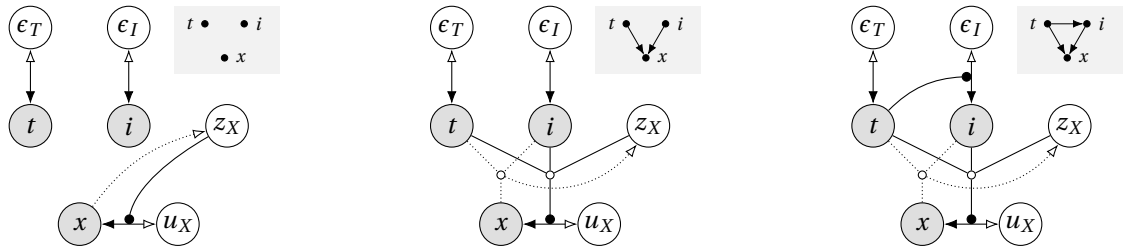
where SetIntensity$(\cdot; i)$ and SetThickness$(\cdot; t)$ refer to the operations that act on an image of a digit and set its intensity to i and thickness to t (details below), x is the resulting image, ϵ^* is the exogenous noise for each variable, and $\sigma(\cdot)$ is the logistic sigmoid. We show some samples in Fig. 7.2.

We use the original MNIST dataset (LeCun et al., 1998) together with the morphometric measurements introduced with Morpho-MNIST (Chapter 6), with new functionality to measure intensity as well as set the intensity and thickness to a given value. We measure the intensity i of an image by following the processing steps in Section 6.2.1 and taking the median value of pixels within the extracted binary mask. Once the intensity is measured, SetIntensity rescales the entire image to match the target intensity, with values clamped between 0 and 255 (images are assumed to be in unsigned 8-bit format). Further, Morpho-MNIST originally proposed only relative thinning and thickening operations. For SetThickness, we extended those operations to absolute values, by calculating the necessary amount of dilation or erosion based on the ratio between target thickness and measured thickness. Finally, we follow Eq. (7.14) to modify each image within the MNIST dataset and randomly split the original training set into a training and validation set.

7.4.2 Experimental setup

We use this setup to study the capabilities of our framework in comparison to models with less causal structure. We adapt the true causal graph from Eq. (7.14) and model thickness and intensity using (conditional) normalising flows and employ a conditional VAE for modelling the image. In particular, we adopt the causal graphs shown in Fig. 7.3 and test a fully independent model (Fig. 7.3a), a conditional decoder model (Fig. 7.3b), as well as our full causal model (Fig. 7.3c). All our experiments were implemented within PyTorch (Paszke et al., 2019) using the Pyro probabilistic programming framework (Bingham et al., 2019).

We use (conditional) normalising flows for all variables apart from the images, which we model using (conditional) deep auto-encoder architectures. The flows consist of components that constrain the support of bounded variables and components relevant for fitting the distribution. We use unit



a Independent **b** Conditional **c** Full

Figure 7.3 Computational graphs of the structural causal models for the Morpho-MNIST example. The image is denoted by x, stroke thickness by t, and image intensity by i. The corresponding causal diagrams are displayed in the top-right corners.

Gaussians as base distributions for all exogenous noise distributions $P(\boldsymbol{\epsilon})$ and, if available, we use the off-the-shelf implementations in PyTorch or Pyro for all transformations. Otherwise, we adapt the available implementations, referring to Durkan et al. (2019) for details. We indicate with θ the modules with learnable parameters.

We model the mechanisms of the thickness t and intensity i as

$$t := f_T(\epsilon_T) = \big(\exp \circ \mathsf{AffineNormalisation} \circ \mathsf{Spline}_\theta\big)(\epsilon_T)\,, \tag{7.15}$$

$$i := f_I(\epsilon_I; t) = \big(\mathsf{AffineNormalisation} \circ \mathsf{sigmoid} \circ \mathsf{ConditionalAffine}_\theta(\hat{t})\big)(\epsilon_I)\,. \tag{7.16}$$

In the independent model, where i is not conditioned on t, we use instead

$$i := f_I(\epsilon_I) = \big(\mathsf{AffineNormalisation} \circ \mathsf{sigmoid} \circ \mathsf{Spline}_\theta \circ \mathsf{Affine}_\theta\big)(\epsilon_I)\,. \tag{7.17}$$

We found that including normalisation layers improves learning dynamics[41] and therefore include flows to perform commonly used normalisation transformations. For doubly bounded variable y we learn the flows in unconstrained space and then constrain them by a sigmoid transform and rescale to the original range using fixed affine transformations with bias $\min(Y)$ and scale $[\max(Y) - \min(Y)]$. We constrain singly bounded values by applying an exponential transform to the unbounded values and using an affine normalisation equivalent to a whitening operation in unbounded log-space. We denote those fixed normalisation transforms as $\mathsf{AffineNormalisation}$ and use a hat to refer to the unconstrained, normalised values (e.g. $\widehat{\mathbf{pa}}_k$). The Spline_θ transformation refers to first-order neural spline flows (Durkan et al., 2019), Affine_θ is an element-wise affine transformation, and $\mathsf{sigmoid}$ refers to the logistic function. $\mathsf{ConditionalAffine}_\theta(\cdot)$ is a regular affine transform whose transformation parameters are predicted by a context neural network taking $\cdot$ as input. In the case of $f_I(\epsilon_I; t)$, the context network is represented by a simple linear transform.

[41] We observed that not normalising the inputs can lead to the deep models prioritising learning the dependence on the variable with largest magnitude. This phenomenon should be investigated further.

Further, we model x using a low-level flow:

$$h_X(u_X; g_X(z_X; \mathbf{pa}_X), \mathbf{pa}_X) = \mathsf{Preprocessing}^{-1}\left(g_X(z_X; \mathbf{pa}_X) + \sigma \cdot u_X\right), \tag{7.18}$$

which reparametrises the noise distribution $P(u_X)$ into another Gaussian distribution with mean given by the high-level mechanism $g_X(z_X; \mathbf{pa}_X)$ and log-variance fixed to $\log\sigma^2 = -5$, and Preprocessing describes a fixed pre-processing transformation.[42]

We implement x's high-level decoder network as a CNN:

$$\begin{aligned} g_X(z_X; \mathbf{pa}_X) = \big(&\mathsf{Conv}_\theta(1;1;1;0) \circ \mathsf{ConvTranspose}_\theta(1;4;2;1) \\ &\circ \mathsf{ReLU} \circ \mathsf{BN}_\theta \circ \mathsf{ConvTranspose}_\theta(64;4;2;1) \circ \mathsf{Reshape}(64,7,7) \\ &\circ \mathsf{ReLU} \circ \mathsf{BN}_\theta \circ \mathsf{Linear}_\theta(1024) \\ &\circ \mathsf{ReLU} \circ \mathsf{BN}_\theta \circ \mathsf{Linear}_\theta(1024)\big)([z_X, \widehat{\mathbf{pa}}_X]), \end{aligned} \tag{7.19}$$

where the operators describe neural network layers as follows: BN is batch normalisation; ReLU the rectified linear activation function; $\mathsf{Conv}(c;k;s;p)$ and $\mathsf{ConvTranspose}(c;k;s;p)$ are a convolution or transposed convolution using a kernel with size k, stride s, padding p, and outputting c channels; $\mathsf{Linear}(h)$ is a linear layer with h output neurons; and Reshape rearranges an array's dimensions to the given shape. Lastly, $[z_X, \mathbf{pa}_X]$ denotes the concatenation of z_X and $\mathbf{pa}_X$, and $z_X \in \mathbb{R}^{16}$.

Equivalently, we implement the the encoder function as a simple CNN that outputs mean and log-variance of a Gaussian with diagonal covariance:

$$\begin{aligned} e_X(x; \mathbf{pa}_X) = \begin{bmatrix} \mathsf{Linear}_\theta(16) \\ \mathsf{Linear}_\theta(16) \end{bmatrix} \Big(\big[\big(&\mathsf{LeakyReLU}(0.1) \circ \mathsf{BN}_\theta \circ \mathsf{Linear}_\theta(100) \circ \mathsf{Reshape}(128 \cdot 7 \cdot 7) \\ &\circ \mathsf{LeakyReLU}(0.1) \circ \mathsf{BN}_\theta \circ \mathsf{Conv}_\theta(128;4;2,1) \\ &\circ \mathsf{LeakyReLU}(0.1) \circ \mathsf{BN}_\theta \circ \mathsf{Conv}_\theta(64;4;2,1)\big)(x),\ \widehat{\mathbf{pa}}_X\big]\Big), \end{aligned} \tag{7.20}$$

where $\mathsf{LeakyReLU}(\ell)$ is the leaky ReLU activation function with a leakiness of ℓ.

We use Adam (Kingma and Ba, 2014) for optimisation with batch size of 256 and a learning rate of 10^{-4} for the encoder–decoder and 0.005 for the covariate flows. We set the number of particles (MC samples) for estimating the ELBO to 4. We use 32 MC samples for estimating reconstruction and counterfactuals (cf. Eq. (7.7)). We train all models for 1000 epochs and report the results of the model with the best validation loss.

[42] We follow the same pre-processing as used with RealNVP (Dinh et al., 2017), which computes $\mathsf{logit}\left(\alpha + (1-\alpha)\frac{x}{256}\right)$, where $\mathsf{logit}(w) = \mathsf{sigmoid}^{-1}(w) = \log\frac{w}{1-w}$.

Table 7.1 Comparison of the associative abilities of the models shown in Fig. 7.3. The image is denoted by x, thickness by t, and intensity by i. Quantities with $\geq$ are lower bounds. MAE refers to the mean absolute error between pixels of the original image and of its reconstruction.

Model	$\log p(x,t,i) \geq$	$\log p(x \mid t,i) \geq$	$\log p(t)$	$\log p(i \mid t)$	$\mathrm{MAE}(x,x')$
Independent	−5925.26	−5919.14	−0.93	−5.19	4.50
Conditional	−5526.50	−5520.37	−0.93	−5.19	4.26
Full	−5692.94	−5687.71	−0.93	−4.30	4.43

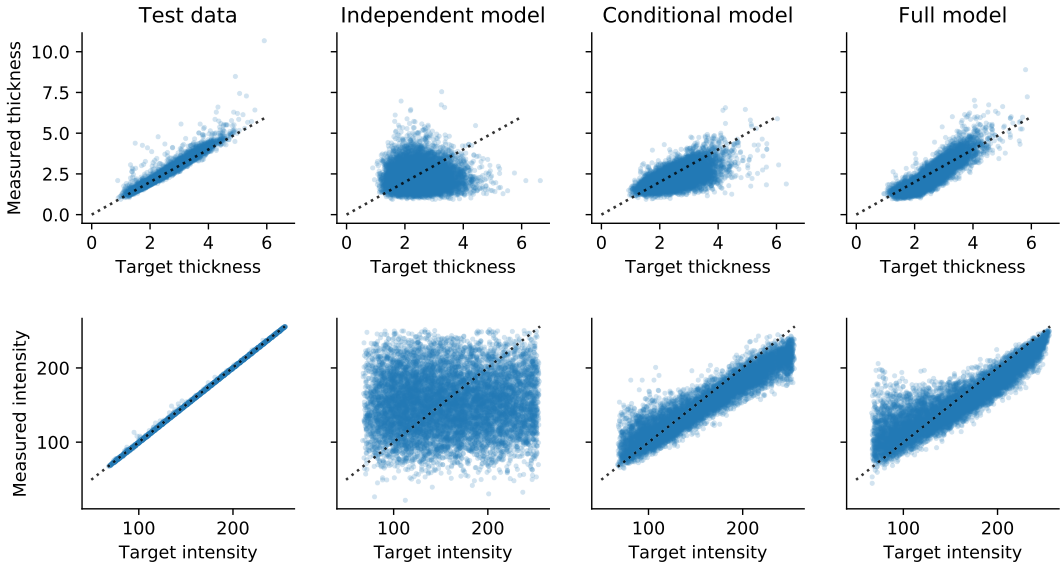
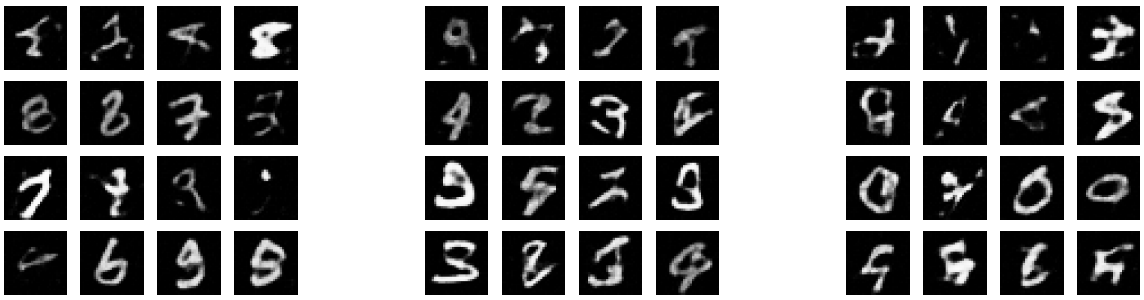


Figure 7.4 Comparison of the target covariates and the corresponding values measured from the generated images. The leftmost column refers to the accuracy of the SetThickness and SetIntensity transforms used in generating the synthetic dataset, and the remaining three columns describe the fidelity of samples generated by each of the learned models. While images sampled from the independent model are trivially inconsistent with the sampled covariates, the conditional and full models show comparable conditioning performance.

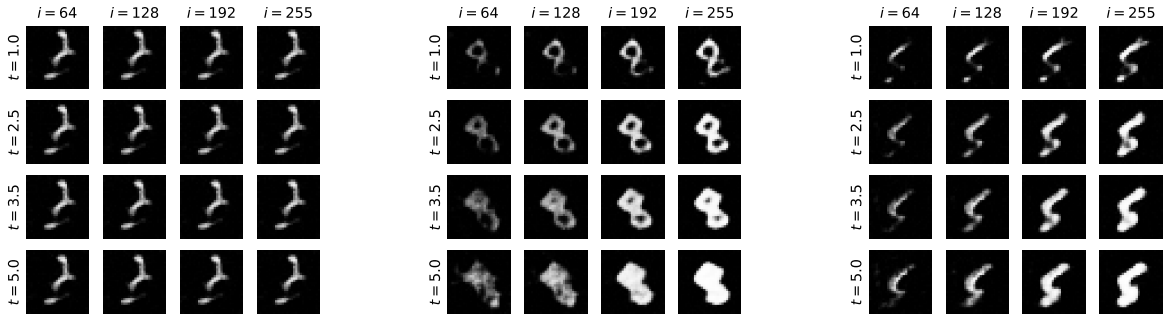
7.4.3 Results

We quantitatively compare the associative capabilities of all models by evaluating their evidence lower bound (Eq. (7.3)), log-likelihoods, and reconstruction errors as shown in Table 7.1. We find that performance improves consistently with the model's capabilities: enabling conditional image generation improves $p(x \mid t,i)$, and adding a causal dependency between t and i improves $p(i \mid t)$. In Fig. 7.4, we visualise how faithfully the generated images correspond to the conditioning covariates, providing more evidence that the image conditioning works as expected. Further, we examine conditional and unconditional image samples in Figs. 7.5 and 7.6, and reconstructions in Fig. 7.7.



a Independent **b** Conditional **c** Full

Figure 7.5 Random samples generated by the independent, conditional and full models. Note how all models appear to have the same unconditional generation capacity.



a Independent **b** Conditional **c** Full

Figure 7.6 Conditional samples generated by the independent, conditional, and full models. The high-level noise, z_X, is shared for all samples from each model, ensuring the same 'style' of the generated digit. The independent model generates images independent of the thickness and intensity values, resulting in identical samples. For the conditional and full models, thickness and intensity change consistently along each column and row, respectively.

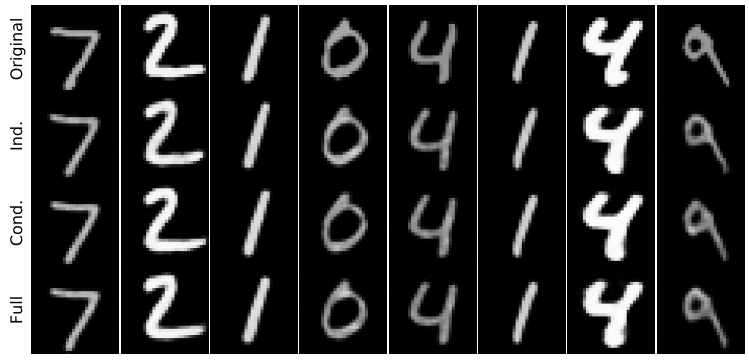


Figure 7.7 Reconstructions. These are computed as Monte Carlo averages approximating $\mathbb{E}_{Q(z_X|e_X(x;\mathbf{pa}_X))}[g_X(z_X;\mathbf{pa}_X)]$, where e_X and g_X are the image encoder and decoder networks. All models seem capable of producing faithful reconstructions.

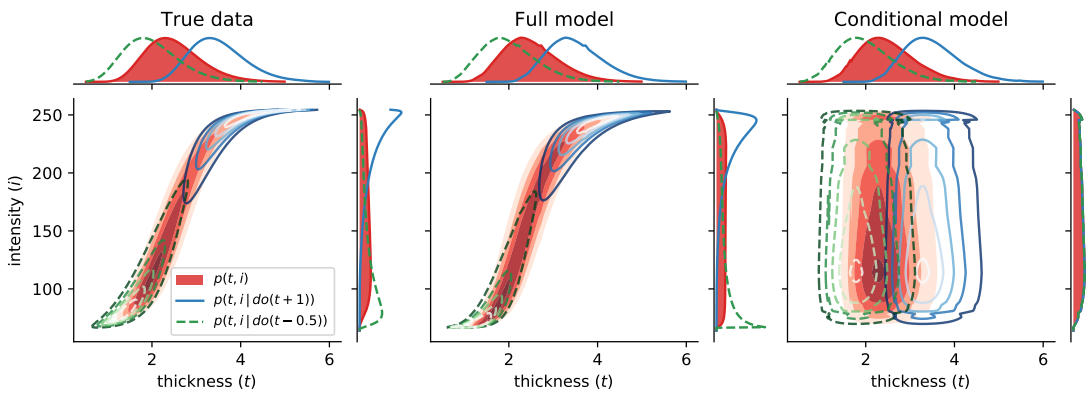


Figure 7.8 Distributions of thickness and intensity in the true data (left), and learned by the full (centre) and conditional (right) models. Contours depict the observational (red, shaded) and interventional joint densities for $\mathrm{do}(t := f_T(\epsilon_T) + 1)$ (blue, solid) and $\mathrm{do}(t := f_T(\epsilon_T) - 0.5)$ (green, dashed).

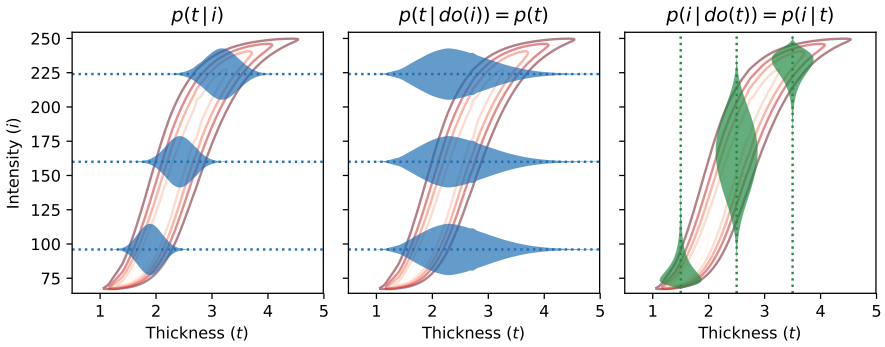


Figure 7.9 Difference between conditioning and intervening, based on the trained full model. The joint density $p(t, i)$ is shown as contours in the background, for reference, and the 'violin' shapes represent the density of one variable when conditioning or intervening on three different values of the other variable. Since t causes i, notice how $p(t \mid i)$ (left) is markedly different from $p(t \mid \mathrm{do}(i))$ (middle), which collapses to $p(t)$. On the other hand, $p(i \mid \mathrm{do}(t))$ and $p(i \mid t)$ (right) are identical.

The interventional distributions can be directly compared to the true generative process. Figure 7.8 shows that the densities predicted by our full model after intervening on t closely resemble the true behaviour. The conditional and independent models operate equivalently and are incapable of modelling the relationship between t and i, capturing only their marginal distributions. As a further illustration of the difference between conditional and interventional distributions, Fig. 7.9 shows how our full model behaves under these different scenarios.

Lastly, we examine the full model's ability to generate counterfactuals. The other two models were omitted as they are incapable of accomplishing interventions, a prerequisite for counterfactual inference. Examples of previously unseen images and generated counterfactuals for broad ranges of interventions are shown in Figs. 7.10 and 7.11. We see that our model is capable of generating convincing counterfactuals that preserve the digit identity while changing thickness and intensity consistently with the underlying causal model.

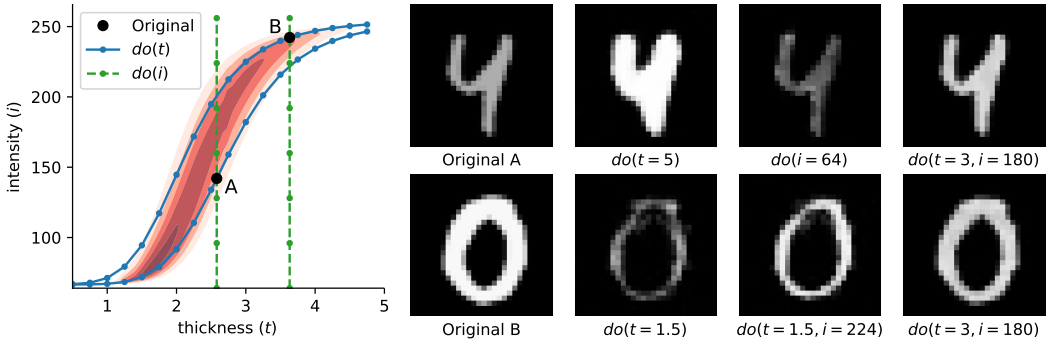


Figure 7.10 Counterfactuals generated by the full model. *Left:* Counterfactual 'trajectories' of two original samples, A and B, as their thickness and intensity are modified, overlaid on the learned joint density $p(t, i)$. *Right:* Original and counterfactual images corresponding to samples A and B.

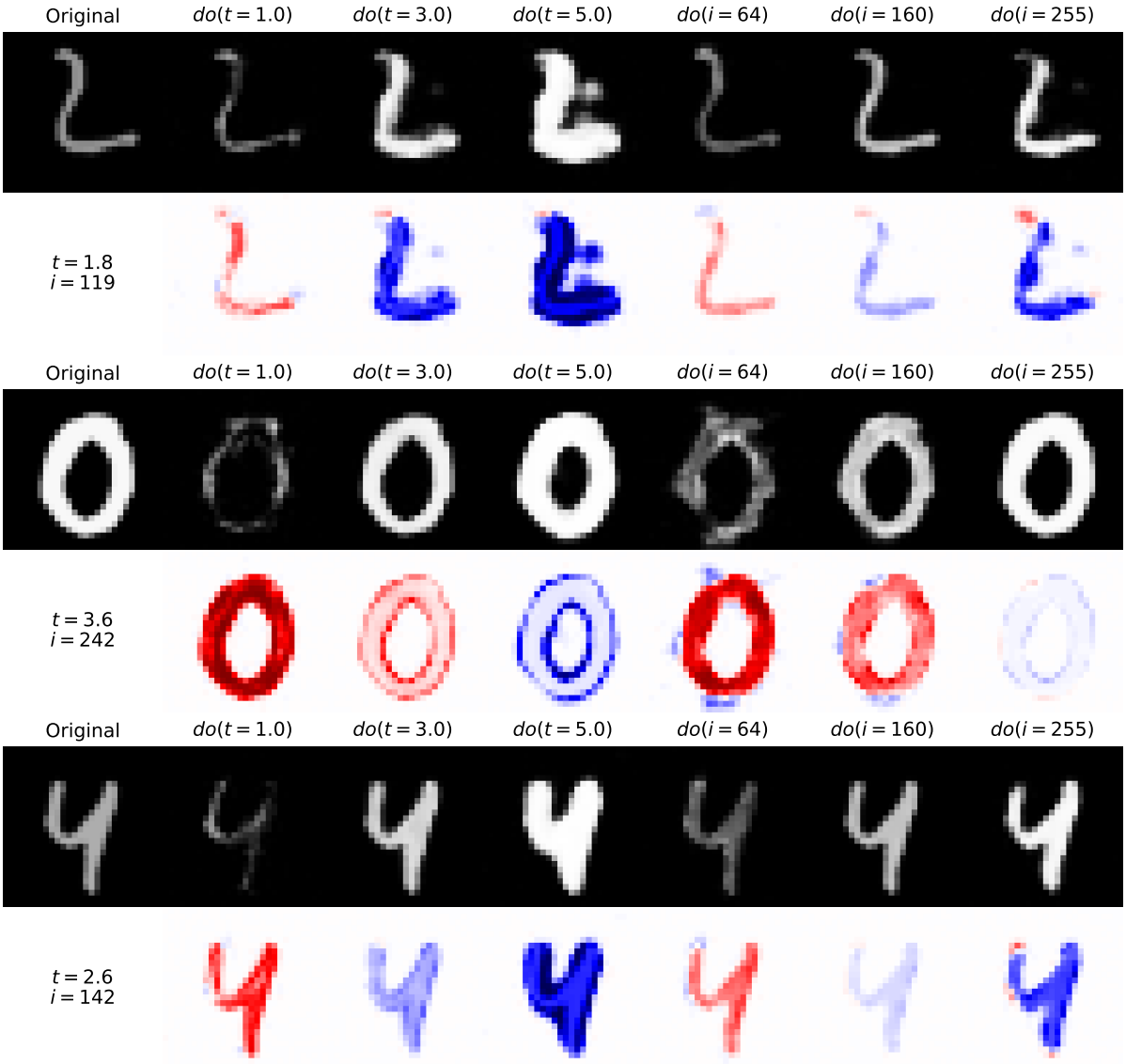


Figure 7.11 Additional counterfactual examples from the full model. The first column shows the original image and true values of the non-imaging data. The even rows show the difference maps between the original image and the corresponding counterfactual image. We observe that all counterfactuals preserve the digits' identity and style. Our model even generates sensible counterfactual images (with some artefacts) in very low-density regions, e.g. '0' with $\mathrm{do}(i = 64)$ (thick but dim), and very far from the original, e.g. '2' with $\mathrm{do}(t = 5.0)$.

7.5 Case study 2: Brain imaging

Our real-world application touches upon fundamental scientific questions in the context of medical imaging: how would a person's anatomy change if particular traits were different? We illustrate with a (simplified) example that our DSCM framework may provide the means to answer such counterfactual queries, which may enable entirely new research into better understanding the physical manifestation of lifestyle, demographics, and disease. Here, we model the appearance of brain MRI scans given the person's age and biological sex, as well as brain and ventricle volumes[43], using population data from the UK Biobank (Sudlow et al., 2015). Figure 7.12 presents several image samples from this dataset. Ventricle and total brain volumes are two quantities that are closely related to brain age (R. Peters, 2006) and can be observed relatively easily. We adopt the causal graph shown in Fig. 7.13a and otherwise follow the same training procedure as for the MNIST experiments.

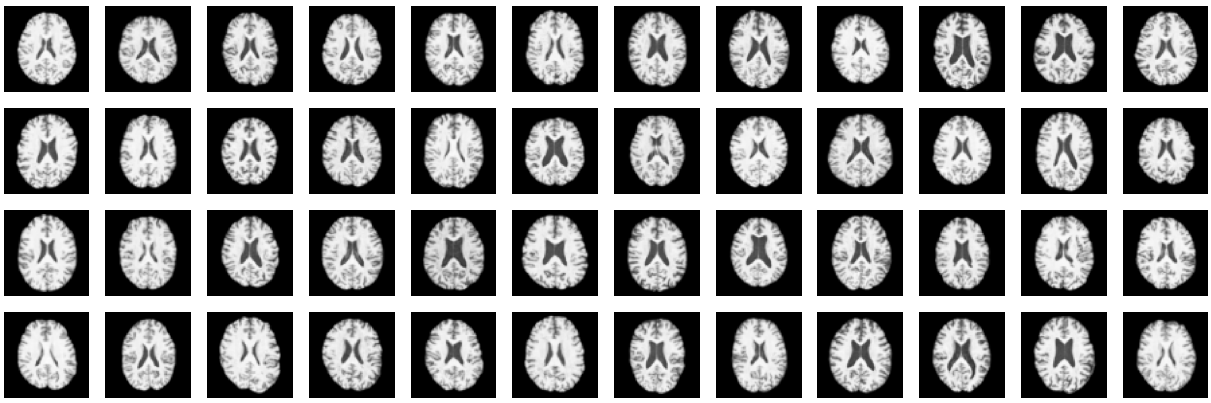
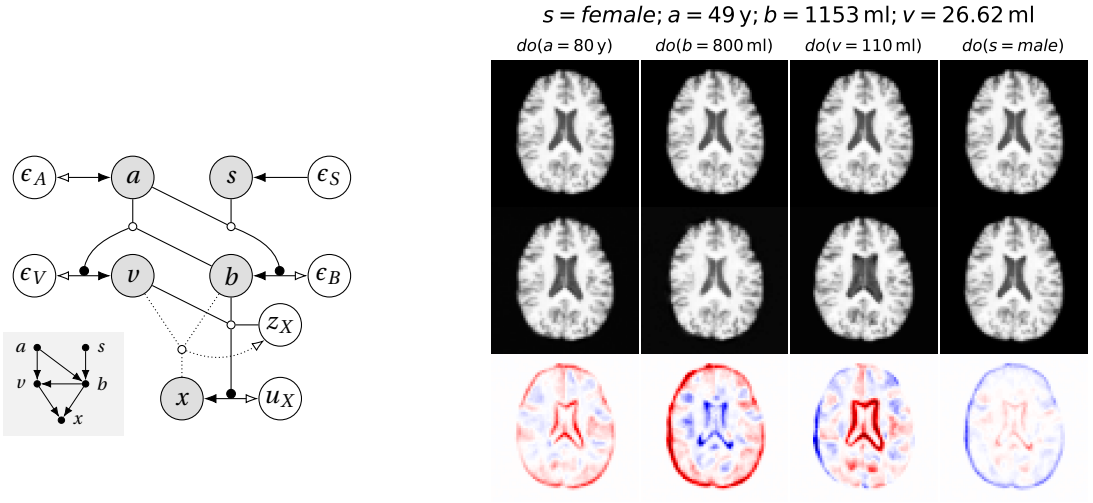


Figure 7.12 Random examplars from the test set of the adopted UK Biobank dataset

7.5.1 Data generation

The original three-dimensional (3D) T1-weighted brain MRI scans were pre-processed by the data providers of the UK Biobank Imaging study using the FSL neuroimaging toolkit (Alfaro-Almagro et al., 2018). This involves skull removal, bias-field correction, and automatic segmentation of brain structures. In addition, we have rigidly registered all scans to the standard MNI atlas space using an in-house image registration tool, which extracted anatomically corresponding mid-axial 2D slices that were used for the reported experiments. The 2D slices were normalised in intensity by mapping the minimum and maximum values inside the brain mask to the range $[0, 255]$. Background pixels outside the brain were set to zero. Age and biological sex for each subject were retrieved from the UK Biobank database along with the pre-computed brain and ventricle volumes. These volumes are derived from the 3D segmentation maps obtained with FSL, and although these are image-derived measurements,

[43] Ventricles are fluid-filled cavities identified as the dark areas in the centre of the brain.



a Computational graph **b** Original image, counterfactuals, and difference maps

Figure 7.13 Brain imaging example. Variables are image (x), age (a), sex (s), and brain (b) and ventricle (v) volumes. The distribution $P(s)$ is Bernoulli, $p(a)$ is an unconditional normalising flow (NF), $p(b \mid a, s)$ and $p(v \mid a, b)$ are conditional NFs, and $p(x \mid b, v)$ is modelled as a conditional VAE. Noise variables are denoted by $\epsilon_\bullet$, and z_X and u_X denote respectively the image x's high- and low-level noise. The counterfactuals (right) show different interventions on the same original brain.

they may serve as reasonable proxies of the true measurements within our (simplified yet plausible) causal model of the physical manifestation of the brain anatomy.

We randomly split the available 13 750 brain images into train, validation and test sets with the respective ratios 70%, 15%, and 15%. During training, we randomly crop the brain slices from their original size of $233 \times 197\,\text{px}^2$ to $192 \times 192\,\text{px}^2$ and use center crops during validation and testing. The cropped images are downsampled by a factor of 3 to a size of $64 \times 64\,\text{px}^2$.

7.5.2 Experimental setup

The setup for the brain imaging experiment closely follows the Morpho-MNIST example. We use the same low-level mechanism for the image x as with MNIST images but change the encoder and decoder functions to a deeper architecture with 5 scales consisting of 3 blocks of $(\mathsf{LeakyReLU}(0.1) \circ \mathsf{BN}_\theta \circ \mathsf{Conv}_\theta)$ each as well as a linear layer that converts to and from the latent space with 100 dimensions. We directly learn the binary probability of the sex s and use the following invertible transforms to model the age a, brain volume b, and ventricle volume v as

$$a := f_A(\epsilon_A) = \big(\exp \circ \mathsf{AffineNormalisation} \circ \mathsf{Spline}_\theta\big)(\epsilon_A)\,, \tag{7.21}$$

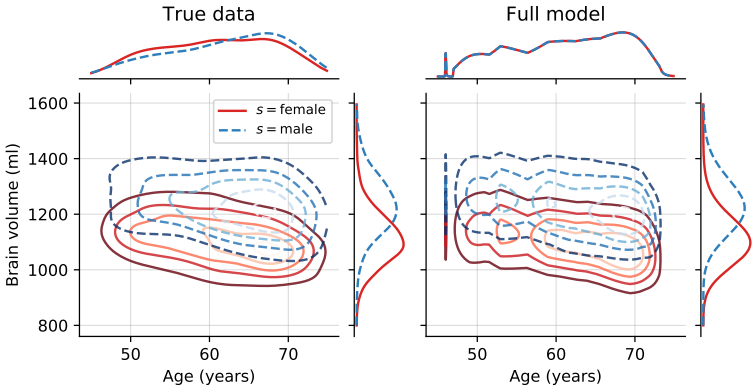
$$b := f_B(\epsilon_B; s, a) = \big(\exp \circ \mathsf{AffineNormalisation} \circ \mathsf{ConditionalAffine}_\theta([s, \widehat{a}])\big)(\epsilon_B)\,, \tag{7.22}$$

$$v := f_V(\epsilon_V; a, b) = \big(\exp \circ \mathsf{AffineNormalisation} \circ \mathsf{ConditionalAffine}_\theta([\hat{b}, \widehat{a}])\big)(\epsilon_V)\,, \tag{7.23}$$

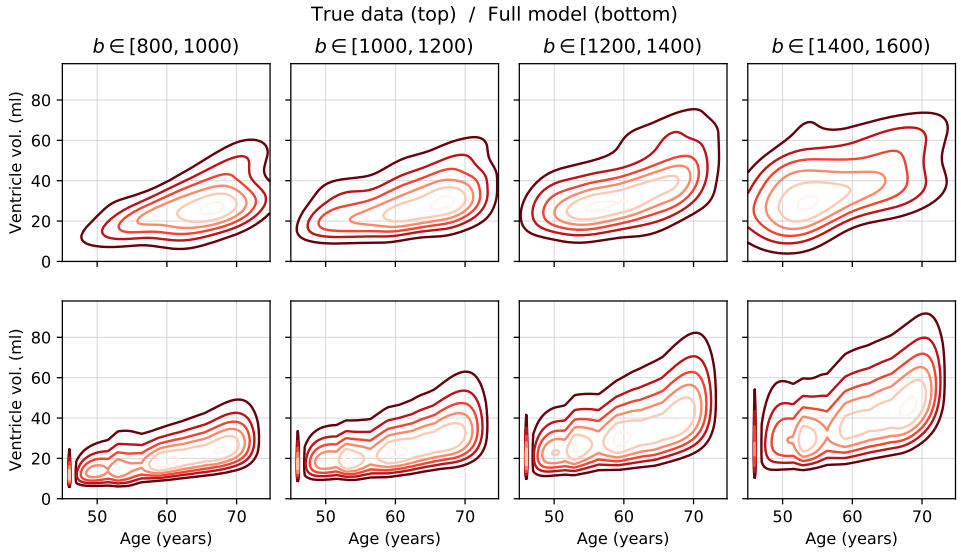
where the context networks are implemented as fully-connected networks with 8 and 16 hidden units, and a LeakyReLU(0.1) nonlinearity.

7.5.3 Results

The learned DSCM is capable of all three levels of the causal hierarchy. We briefly present the analysis of the observational level in Fig. 7.14 (for covariates) and Figs. 7.15 to 7.17 (for images). Here we focus on counterfactuals, shown in Fig. 7.13b with more detailed examples in Fig. 7.18. The difference maps show plausible counterfactual changes: increasing age causes slightly larger ventricles while decreasing the overall brain volume (first column). In contrast, directly changing brain volume has an opposite effect on the ventricles compared to changing age (second column). Intervening on ventricle volume has a much more localised effect (third column), while intervening on the categorical variable of biological sex has smaller yet more diffuse effects. Note how the anatomical 'identity' (such as the cortical folding) is well preserved after each intervention.



a Age vs. brain volume: $p(a, b \mid s)$. Here we see differences in head size across biological sexes (reflected in brain volume), as well as a downward trend in brain volume as age progresses.



b Age vs. ventricle volume: $p(a, v \mid b \in \cdot)$. As expected from the literature (R. Peters, 2006), we observe a consistent increase in ventricle volume with age, in addition to a proportionality relationship with the overall brain volume.

Figure 7.14 Densities for the true data (KDE) and for the learned model. The overall trends and interactions present in the true data distribution seem faithfully captured by the model.

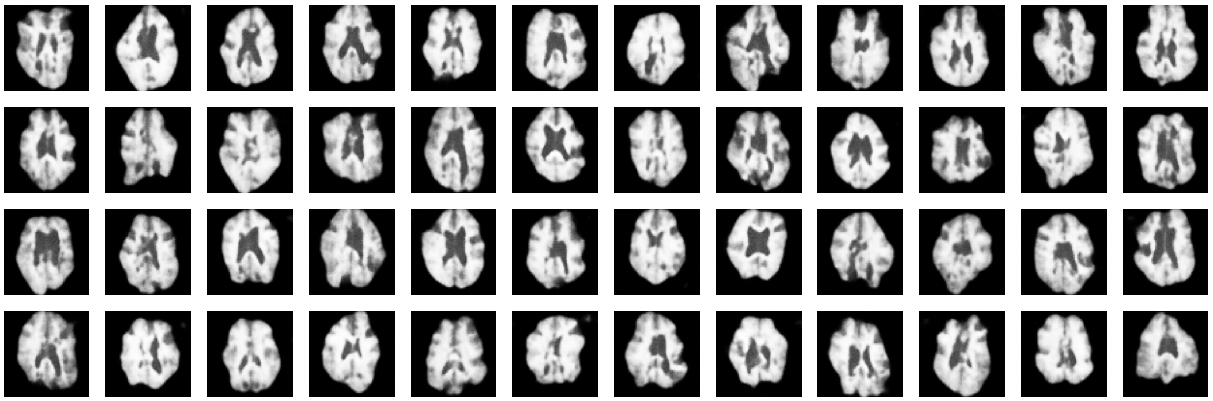


Figure 7.15 Random samples from the model trained on the UK Biobank dataset

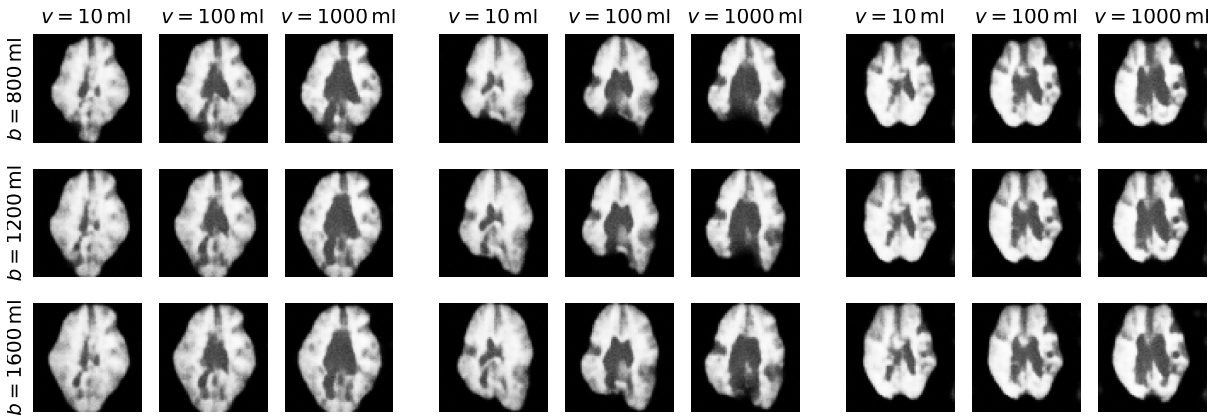


Figure 7.16 Conditional samples from the model trained on the UK Biobank dataset. Images in each 3×3 block share the same high-level noise vector, z_X. Each row consistently changes the brain size, whereas each column changes the ventricle volume.

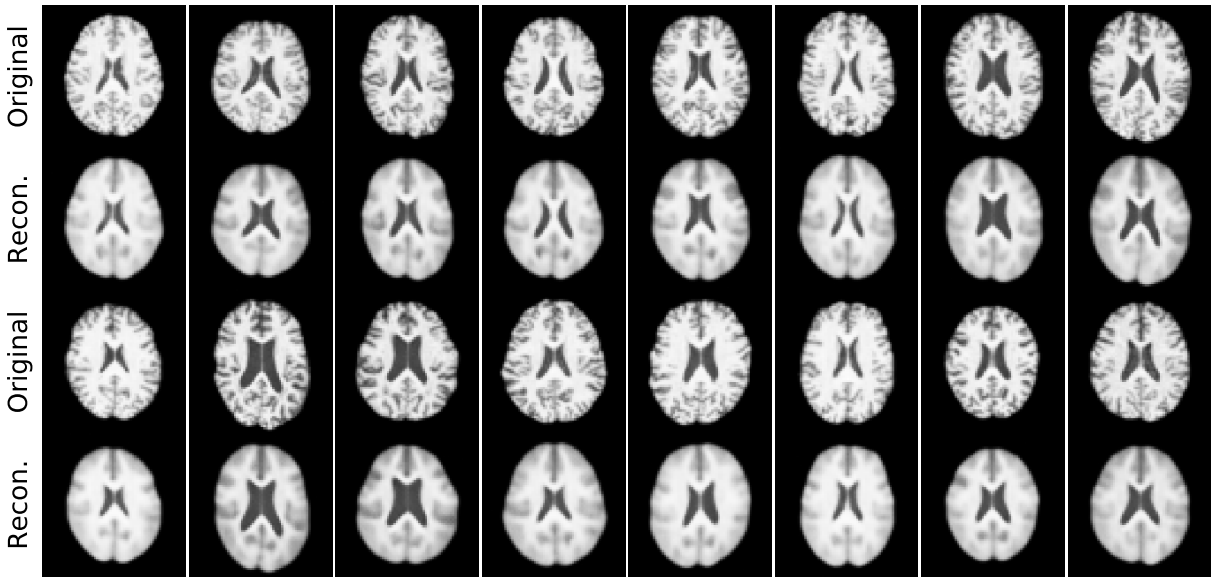


Figure 7.17 Original samples and reconstructions from the model trained on the UK Biobank dataset

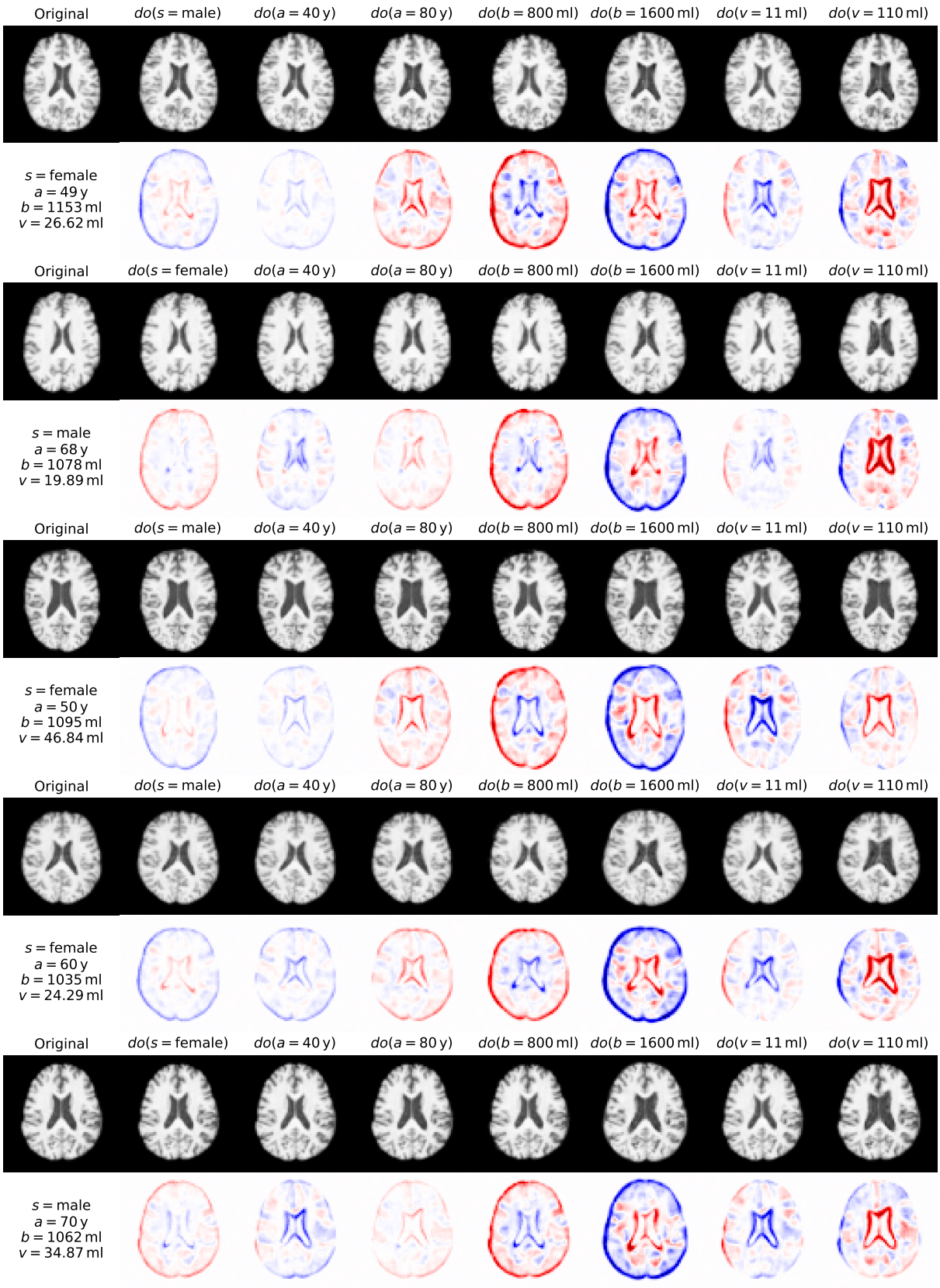


Figure 7.18 Original samples and counterfactuals from the model trained on the UK Biobank dataset. The first column shows the original image and true values of the non-imaging data. The even rows show the difference maps between the original image and the corresponding counterfactual image.

7.6 Conclusion

We introduce a novel general framework for fitting SCMs with deep mechanisms. Our deep SCM (DSCM) framework fulfils all three rungs of Pearl's ladder of causation—in particular, it is the first to enable efficient abduction of exogenous noise, permitting principled counterfactual inference. We demonstrate the potential of DSCMs with two case studies: a synthetic task of modelling Morpho-MNIST digits with a known causal structure and a real-world example with brain MRI.

The ability to correctly generate plausible counterfactuals could greatly benefit a wide variety of possible applications, e.g.: *explainability*, where differences between observed and counterfactual data can suggest causal explanations of outcomes; *data augmentation*, as counterfactuals can extrapolate beyond the range of observed data (e.g. novel combinations of attributes); and *domain adaptation*, since including the source of the data as an indicator variable in the causal model could enable generating counterfactual examples in a relevant target domain.

The proposed method does not come without limitations to be investigated in future work. Like the related approaches, the current setup requires all variables to be observed when computing a counterfactual, which may limit its applicability in certain scenarios. This could be alleviated by imputing the missing data via MCMC or learning auxiliary distributions. Further work should study more closely the dynamic behaviour of deep mechanisms in SCMs during training. While not observed in our experiments, neural networks may not learn to cleanly separate the roles of its inputs on the output as expected—which could require custom counterfactual regularisation similar to losses used in image-to-image translation (Xia et al., 2019) and explainability work (Singla et al., 2020). The use of such flexible models also raises questions about the identifiability of the 'true' mechanism, as counterfactuals may not be uniquely defined. Lastly, it would be interesting to examine whether this framework can be applied to causal discovery, attempting to uncover plausible causal structures from data.

Chapter 8

Conclusion

8.1 Summary of achievements

Recall from Section 1.2 that the aims of my PhD research have been

(A) **To establish the utility of causality in imaging problems;** and

(B) **To develop effective methodologies for jointly modelling images and non-imaging data.**

In what follows, I summarise how the main conclusions presented in each chapter make concrete progress towards these goals, and list pointers to some of the impact these contributions have already made since publication.

As seen in **Chapter 3**, we have successfully applied the framework of causal reasoning to typical medical image analysis scenarios. First, this revealed that the suitability of the most common strategies for mitigating data scarcity (viz. semi-supervision and data augmentation) may depend on the causal direction of the task. Second, it enabled us to systematically categorise various familiar forms of data mismatch in medical imaging—which introduce distinct biases and warrant different solutions—based on the causal structure of the data-generating process. The reviewed literature further surfaced other important challenges such as missingness and measurement bias, which can also be studied under a causal interpretation. Finally, we saw how these causal considerations relate to and complement recent best practices for reporting machine-learning studies in healthcare.

Chapter 4 determined that a certain class of nonparametric Bayesian model, Dirichlet process Gaussian mixture models (DPGMMs), is useful for modelling image intensity distributions. By applying them to single- and multi-modality MRI intensities, we verified that DPGMMs are powerful tools for representing multidimensional probability densities, as their capacity can adapt to the complexity

of the data. In addition, the DPGMMs' tractable density function led to a practical algorithm for smoothly matching intensity distributions between images, which we termed nonparametric density flows (NDFLOW). Like the competing state-of-the-art method, this normalisation made brain tissue intensities more consistent across images from all sites and appeared to remove some scanner biases, reducing the effects of acquisition shift (whose causal interpretation is presented in Chapter 3). NDFLOW was used by Yao et al. (2019) in a spectral imaging application, specifically for matching spectral channels to RGB when reconstructing depth maps.

Building on the intuitions developed in Chapter 4, **Chapter 5** employed the same kind of nonparametric mixtures in a (benign) face recognition application. The full model is capable of grouping face embeddings and sparse name labels under an unbounded number of distinct identities, whose co-occurrence patterns are represented by context labels. The extended model further integrates environmental signals as contextual side-information, all in a unified probabilistic graphical model. This illustrates not only that these mixture models can tractably represent joint distributions over multiple variables (here, image-derived features and non-imaging covariates), but also that their nonparametric nature allows them to seamlessly scale up as novel combinations of values are observed. Grayson et al. (2020) showcased the larger system that includes this identity recognition module, along with its other functionalities such as awareness of gaze and pose of individuals surrounding the user and a tailored interface with audio feedback.

To facilitate research into models and representations of heterogeneous data, **Chapter 6** introduced an experimental sandbox based on images of handwritten digits, including measurable shape attributes and controllable image perturbations. The proposed morphometric approach to quantitative evaluation enables the unprecedented analysis of generative behaviour. Specifically, it can determine to what extent a generative model of images reproduces the marginal distribution of shape attributes, and also whether its latent representation has disentangled dimensions corresponding to specific factors of shape variation. For conditional image models, it can measure how faithfully the shapes of sampled images match the conditioning inputs. This toolkit was later extended for the work in Chapter 7 with image transformations that force an attribute to a given value (e.g. 'SetThickness'), enabling the design of experiments using synthetic data with a prescribed causal structure.

Since publication, this framework has already been adopted in several studies. For example, Prabhu (2019) introduced an MNIST-like dataset of handwritten characters in the Kannada language,[44] and used Morpho-MNIST to compare its shape statistics against the original MNIST. Georgopoulos et al. (2020) and Lotfollahi et al. (2019) evaluate whether their conditional image generation methods trained on perturbed Morpho-MNIST images can handle unseen combinations of attributes. Pati

[44] This new dataset features in a dedicated Kaggle competition: https://www.kaggle.com/c/Kannada-MNIST.

and Lerch (2020) use our MNIST shape attributes in experiments with their proposed auto-encoder regularisation method, analysing various disentanglement and reconstruction metrics. In a creative use of the proposed perturbations, Le Zhang*, Tanno* et al. (2020) investigate uncertainty of manual image segmentations by simulating different types of annotator (under-segmentation: thinned; over-segmentation: thickened; wrong segmentation: fractured and thickened). Lastly, Robinson et al. (2020) validate a novel domain adaptation technique on data with synthetic domain shift, harmonising sets of thin digits toward thickened and/or slanted digits.

Addressing both aims stated above, **Chapter 7** finally presented the framework of deep structural causal models (DSCMs), which bridges causality theory and deep probabilistic modelling. When sufficient domain knowledge is available, SCMs are a natural way to formulate causally informed models of the data-generating process. Further, the implementation of causal mechanisms as deep neural networks enables the inclusion of high-dimensional variables, such as images, into models with generative, interventional, and counterfactual capabilities. By adopting a theoretically motivated approach to counterfactual inference, DSCMs are able to output plausible images in response to interventions on arbitrary subsets of covariates, while preserving fine details of the original image. Though not experimented with in this work, our proposal of using adversarially trained mechanisms was since validated by Dash and Sharma (2020), who analysed DSCMs both for counterfactual image generation and for assessing fairness of an image classifier.

8.2 Limitations, open questions, and outlook

In this final section, we reflect on some issues that remained unaddressed and open challenges that warrant further research.

8.2.1 Validity of causal models

One of the greatest limitations of this work—and indeed of causality research more generally—is the need to rely on arguably strong causal assumptions about the data-generating process. When the goal is estimating causal effects (level two of the hierarchy, Section 2.3.1), *do*-calculus is able to identify how the query can be computed from data, or detect whether that is impossible without additional assumptions (Pearl, 2009). Also, when observational and experimental data is available, some level-one and level-two assumptions can be tested against the evidence and potentially falsified. However, when aiming for counterfactual prediction, concerns about *identifiability* are even more important. As there may be (infinitely) many different models that entail the same observational and interventional distributions, one cannot generally define a unique counterfactual (J. Peters et al., 2017,

Sec. 6.8). The use of 'black-box' components as causal mechanisms may further exacerbate these issues, already present even for simple linear models. Further research may be needed to develop theoretical guarantees for level-three models, especially when applied to high-dimensional data.

Moreover, as mentioned in Section 2.3.1 and Section 7.6, true evaluation of counterfactual predictions is impossible, as counterfactual outcomes by definition can never be observed. With controlled environments, however, one may simulate certain kinds of interventions and compare the result with the model predictions. In a preliminary exploration of this idea, we generated synthetic counterfactual Morpho-MNIST images with do(thickness + 2), and computed the pixel-level mean absolute errors: the full model achieved 17.6, while the conditional and independent models reached 41.6 and 31.8, respectively. Alternatively, in constrained settings, it may be possible to compare counterfactual predictions against paired data. For example, assuming our neuroimaging graph from Section 7.5 is correct (i.e. ignoring possible unobserved confounders), we may use a second MRI scan of a brain a few years later as an approximate counterfactual on age. In other applications, it may be possible to leverage symmetries, e.g. comparing left and right organs or limbs of the same individual.

8.2.2 Beyond fully observed, fully specified causal models

Another great challenge in causal inference and causal machine learning is the presence of *unobserved confounders*, i.e. when the direct causal relationships between observed variables do not account for all dependences. At present, the DSCM framework introduced in Chapter 7 is applicable only to causal models wherein all latent variables are assumed independent and each affects a single observed variable. This is what permits the proposed efficient abduction procedure, since the posterior distribution of noise given observations factorises over all mechanisms (Eq. (7.5)). Addressing the case of unobserved variables could significantly broaden the scope of DSCMs to more realistic applications. Preliminary ideas involve training recognition models for each latent variable given its Markov blanket, as done by C. Zhang et al. (2020).

A related issue that was briefly mentioned in Section 3.5 is *missingness*. In other words, certain variables may not be observed for all units (e.g. labels in a semi-supervised scenario), and whether a value is missing may be dependent on other variables. This is a pervasive obstacle in real-life applications, especially in healthcare (Ghassemi et al., 2019; N. D. Shah et al., 2018). It could therefore be valuable to combine causal interpretations and theoretical results regarding this phenomenon (Daniel et al., 2012; Mohan et al., 2013) with powerful computational models such as DSCMs to extend our predictive capabilities in such challenging environments.

Lastly, in complex domains like healthcare, where data is notoriously noisy (S. Hoffman and Podgurski, 2013; N. D. Shah et al., 2018), it would be beneficial to integrate knowledge of *measurement*

bias mechanisms (Hernán and Cole, 2009; Shahar, 2009) into our causal models. We discussed in Section 3.4.1 a special case of differential measurement bias wherein the domain affects image acquisition (i.e. acquisition shift). However, more general mechanisms may affect the measurement of other variables, and it would be useful to explore how such considerations can be incorporated into practical algorithm design for high-dimensional data.

8.2.3 Causal understanding of high-dimensional features

Most of causality theory has originally been developed in the contexts of epidemiology, econometrics, social sciences, and other fields involving applied statistics, wherein the variables of interest tend to be scalars. In this case, it is easy to reason about the meanings of interventions and causal effects, as the values can only increase or decrease, or e.g. flip between 0 and 1.

The situation is considerably more nuanced when high-dimensional variables are involved, as interventions can have qualitatively distinct behaviours and may be understood as acting on *high-level features* rather than purely on the raw data. For example, how do we formulate an image intervention such as 'shift the object to the left by 3 px'? How could this expression be distinguished from 'change the object's colour to red'? How would we even express the concept of an 'object' in the image? Would a thickening transformation on an MNIST digit be considered an intervention on the image or on the original (anticausal) thickness? What about a change in the data domain, such as resampling an image to a different resolution or a change in coordinate system (e.g. during spatial registration)? Could normalising the intensities of a medical scan be interpreted as an intervention neutralising the effect of scanner contrast?

Chalupka et al. (2017) and Rubenstein et al. (2017) have developed theoretical foundations surrounding micro- and macro-scale causal phenomena that may help frame such concepts and queries, while relevant methodologies regarding causal manipulation of image features have also been proposed for specific applications (e.g. data augmentation, Ilse et al., 2020; adversarial robustness, C. Zhang et al., 2020). I believe the endeavour of causal modelling with heterogeneous data would greatly benefit from a comprehensive framework that is able to represent relationships of high-level image attributes not only with non-imaging covariates, but also with image-derived annotations, such as segmentation masks or object bounding boxes. Such a fine-grained representation of the problem domain could, for example, lead to the design of disentangled model architectures with task-relevant inductive biases, enabling reasoning about specifically targeted interventions. We touch upon broader impacts in the final section, below.

8.2.4 Practical implications and applications

Lastly, it is crucially important to discuss the practical impact of developing this causal machinery for high-dimensional, heterogeneous data. Why does it matter that we are able to develop conceptual and computational models of complex causal systems?

As argued at length in Chapter 3, one of the key benefits of reasoning causally about a problem domain is *transparency*, by offering a clear and precise language to communicate assumptions about the collected data. In a similar vein, models whose architecture mirrors an assumed causal graph can be desirable in applications where *interpretability* is important (Moraffah et al., 2020). Furthermore, more powerful methods for counterfactual predictions would also contribute to progress in *explainability* of black-box models.

Another promising direction of investigation is *transferability*. Recall that the principle of independent causal mechanisms (Section 2.3.2) postulates that each variable in a system is generated by its own autonomous process, and that this is the basis for the arguments about the stability of certain mechanisms in different environments under data mismatch (Section 3.4). Especially accounting for the considerations in Section 8.2.3, framing existing domain adaptation and generalisation methods under a causal perspective could provide novel insights and perhaps enable novel techniques for improving external validity. Moreover, there would be merit in studying the reuse of learned causal mechanisms in transfer learning and multi-task learning scenarios, potentially leading to more data- and resource-efficient methods by training specialised sub-models.

As a final comment, I would like to draw attention to the complex but vital issue of *fairness*. There is a growing awareness that certain groups of individuals are systematically and disproportionately harmed by ramifications of data-driven technologies, which reproduce and reinforce societal biases. The healthcare context is no exception (Grote and Berens, 2020). Considerable research efforts are therefore being devoted to developing fair ML methodologies (Barocas et al., 2019), not only for building fair models, but also for auditing existing systems. In particular, causal reasoning has proven to be an effective language for representing and operationalising certain notions of fairness (Chiappa and Isaac, 2019; Kilbertus et al., 2017; Kusner et al., 2017; Barocas et al., 2019, Ch. 4). Leveraging more sophisticated models of causality for high-dimensional, heterogeneous data, such as the extensions mentioned in Section 8.2.3, it may be possible to apply such causal fairness frameworks to the medical imaging domain. Given the considerations outlined above—on transparency of assumptions, model interpretation, explanation of predictions, robustness across environments and tasks, and fairness—it is evident that a causal perspective on machine learning contributes to the goal of building systems that are not just highly performant, but crucially also safe and reliable for all stakeholders.

Bibliography

Academy of Medical Royal Colleges (Jan. 2019). *Artificial Intelligence in Healthcare*. London, UK. URL: https://www.aomrc.org.uk/reports-guidance/artificial-intelligence-in-healthcare (cit. on p. 59).

Priyanka Agrawal, Lavanya Sita Tekumalla and Indrajit Bhattacharya (2013). 'Nested hierarchical Dirichlet process for nonparametric entity-topic analysis'. In: *Machine Learning and Knowledge Discovery in Databases – ECML PKDD 2013*. Vol. 8189. LNCS. Springer, Berlin, Heidelberg, pp. 564–579. DOI: 10.1007/978-3-642-40991-2_36 (cit. on p. 99).

David J. Aldous (1985). 'Exchangeability and related topics'. In: *École d'Été de Probabilités de Saint-Flour XIII – 1983*. Ed. by P. L. Hennequin. Vol. 1117. Lecture Notes in Mathematics. Springer Berlin Heidelberg. DOI: 10.1007/BFb0099421 (cit. on pp. 24, 25).

Fidel Alfaro-Almagro, Mark Jenkinson, Neal K. Bangerter, Jesper L. R. Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N. Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee et al. (2018). 'Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank'. *NeuroImage* 166, pp. 400–424 (cit. on p. 144).

Brandon Amos, Bartosz Ludwiczuk and Mahadev Satyanarayanan (2016). *OpenFace: A general-purpose face recognition library with mobile applications*. Tech. rep. CMU-CS-16-118. CMU School of Computer Science (cit. on pp. 89, 97).

Dragomir Anguelov, Kuang-chih Lee, Salih B. Gökturk and Baris Sumengen (2007). 'Contextual identity recognition in personal photo albums'. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2007)*, pp. 1–7. DOI: 10.1109/CVPR.2007.383057 (cit. on p. 86).

Grigory Antipov, Moez Baccouche and Jean-Luc Dugelay (2017). 'Face aging with conditional generative adversarial networks'. In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 2089–2093. DOI: 10.1109/ICIP.2017.8296650 (cit. on p. 136).

Charles E. Antoniak (1974). 'Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems'. *The Annals of Statistics* 2.6, pp. 1152–1174. JSTOR: 2958336 (cit. on pp. 28, 64, 89).

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani and David Lopez-Paz (2019). *Invariant risk minimization*. arXiv: 1907.02893 (cit. on p. 55).

Sanjeev Arora, Andrej Risteski and Yi Zhang (2018). 'Do GANs learn the distribution? Some theory and empirics'. In: *International Conference on Learning Representations (ICLR 2018)* (cit. on p. 108).

John Ashburner and Karl J. Friston (2005). 'Unified segmentation'. *NeuroImage* 26.3, pp. 839–851. DOI: 10.1016/j.neuroimage.2005.02.018 (cit. on pp. 22, 64).

Mathieu Aubry, Daniel Maturana, Alexei A. Efros, Bryan C. Russell and Josef Sivic (2014). 'Seeing 3D chairs: exemplar part-based 2D–3D alignment using a large dataset of CAD models'. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*. IEEE, pp. 3762–3769. Dataset URL: https://www.di.ens.fr/willow/research/seeing3Dchairs/ (cit. on p. 107).

Maria-Florina Balcan and Avrim Blum (2006). 'An augmented PAC model for semi-supervised learning'. In: Chapelle, Schölkopf and Zien, 2006. Chap. 22, pp. 397–419 (cit. on p. 51).

Elias Bareinboim, Juan D. Correa, Duligur Ibelind and Thomas Icard (2020). 'On Pearl's hierarchy and the foundations of causal inference'. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. Ed. by Rina Dechter, Hector Geffner and Joseph Y. Halpern. Forthcoming. ACM Books. eprint: https://causalai.net/r60.pdf (cit. on p. 37).

Elias Bareinboim and Judea Pearl (2016). 'Causal inference and the data-fusion problem'. *Proceedings of the National Academy of Sciences of the United States of America* 113.27, pp. 7345–7352. DOI: 10.1073/pnas.1510507113 (cit. on pp. 14, 47, 53, 57).

Elias Bareinboim, Jin Tian and Judea Pearl (2014). 'Recovering from selection bias in causal and statistical inference'. In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI 2014)*. AAAI, pp. 2410–2416 (cit. on p. 57).

Solon Barocas, Moritz Hardt and Arvind Narayanan (2019). *Fairness and Machine Learning*. URL: https://fairmlbook.org/ (cit. on p. 156).

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sebastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal and Christopher Pal (2020). 'A meta-transfer objective for learning to disentangle causal mechanisms'. In: *International Conference on Learning Representations (ICLR 2020)* (cit. on p. 128).

Jan-Philip Bergeest and Florian Jäger (2008). 'A comparison of five methods for signal intensity standardization in MRI'. In: *Bildverarbeitung für die Medizin 2008*. Informatik aktuell. Springer, Berlin, Heidelberg, pp. 36–40. DOI: 10.1007/978-3-540-78640-5_8 (cit. on p. 71).

José M. Bernardo and Adrian F. M. Smith (2000). *Bayesian Theory*. Wiley Series in Probability and Statistics. Chichester, US: John Wiley & Sons, Inc. DOI: 10.1002/9780470316870 (cit. on p. 65).

Alejandro Betancourt, Pietro Morerio, Carlo S. Regazzoni and Matthias Rauterberg (2015). 'The evolution of first person vision methods: a survey'. *IEEE Transactions on Circuits and Systems for Video Technology* 25.5, pp. 744–760. DOI: 10.1109/TCSVT.2015.2409731 (cit. on p. 86).

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall and Noah D. Goodman (2019). 'Pyro: deep universal probabilistic programming'. *Journal of Machine Learning Research* 20.28 (cit. on p. 137).

Mikołaj Bińkowski, Dougal Sutherland, Michael Arbel and Arthur Gretton (2018). 'Demystifying MMD GANs'. In: *International Conference on Learning Representations (ICLR 2018)*. arXiv: 1801.01401 (cit. on p. 108).

Christopher M. Bishop (2006). *Pattern Recognition and Machine Learning*. New York: Springer (cit. on pp. 30, 31).

David Blackwell and James B. MacQueen (1973). 'Ferguson distributions via Pólya urn schemes'. *The Annals of Statistics* 1.2, pp. 353–355. JSTOR: 2958020 (cit. on p. 24).

David M. Blei, Thomas L. Griffiths and Michael I. Jordan (2010). 'The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies'. *Journal of the ACM* 57.2. DOI: 10.1145/1667053.1667056 (cit. on p. 99).

David M. Blei and Michael I. Jordan (2006). 'Variational inference for Dirichlet process mixtures'. *Bayesian Analysis* 1.1, pp. 121–144. DOI: 10.1214/06-BA104 (cit. on p. 65).

Patrick Blöbaum, Shohei Shimizu and Takashi Washio (2015). 'Discriminative and generative models in causal and anticausal settings'. In: *Advanced Methodologies for Bayesian Networks (AMBN 2015)*. Vol. 9505. LNCS. Springer, Cham, pp. 209–221. DOI: 10.1007/978-3-319-28379-1_15 (cit. on pp. 48, 51).

Harry Blum (1967). 'A transformation for extracting new descriptors of shape'. In: *Models for the perception of speech and visual form*. MIT Press, pp. 362–380 (cit. on pp. 109, 110).

Wacha Bounliphone, Eugene Belilovsky, Matthew B. Blaschko, Ioannis Antonoglou and Arthur Gretton (2016). 'A test of relative similarity for model selection in generative models'. In: *International Conference on Learning Representations (ICLR 2016)*. arXiv: 1511.04581 (cit. on p. 116).

Charles Bouveyron and Stéphane Girard (2009). 'Robust supervised classification with mixture models: learning from data with uncertain labels'. *Pattern Recognition* 42.11, pp. 2649–2658. DOI: 10.1016/j.patcog.2009.03.027 (cit. on p. 90).

Yaroslav Bulatov (Sept. 2011). *notMNIST dataset*. URL: https://yaroslavvb.blogspot.co.uk/2011/09/notmnist-dataset.html (visited on 08/05/2018) (cit. on p. 107).

Mariano Cabezas, Arnau Oliver, Xavier Lladó, Jordi Freixenet and Meritxell Bach Cuadra (2011). 'A review of atlas-based segmentation for magnetic resonance brain images'. *Computer Methods and Programs in Biomedicine* 104.3, e158–e177. DOI: 10.1016/j.cmpb.2011.07.015 (cit. on p. 64).

Daniel C. Castro (2016). 'Dirichlet process mixture models applied to brain image segmentation'. MRes thesis. London: Imperial College London (cit. on pp. 16, 19, 63).

Daniel C. Castro and Ben Glocker (2016). 'Dirichlet process mixture models: application to brain image segmentation'. In: *NIPS 2016 Workshop on Practical Bayesian Nonparametrics* (Barcelona, Spain). HDL: 10044/1/65183 (cit. on pp. 16, 63).

Daniel C. Castro and Ben Glocker (2018). 'Nonparametric density flows for MRI intensity normalisation'. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018* (Granada, Spain). Vol. 11070. LNCS. Springer, Cham, pp. 206–214. DOI: 10.1007/978-3-030-00928-1_24. arXiv: 1806.02613 (cit. on pp. 16, 63).

Daniel C. Castro and Sebastian Nowozin (2018a). 'Contextual face recognition with a nested-hierarchical nonparametric identity model'. In: *NeurIPS 2018 Workshop on All of Bayesian Nonparametrics* (Montréal, Canada). arXiv: 1811.07753 (cit. on pp. 16, 83).

Daniel C. Castro and Sebastian Nowozin (2018b). 'From face recognition to models of identity: a Bayesian approach to learning about unknown identities from unsupervised data'. In: *Computer Vision – ECCV 2018* (Munich, Germany). Vol. 11206. LNCS. Springer, Cham, pp. 764–780. DOI: 10.1007/978-3-030-01216-8_46. arXiv: 1807.07872 (cit. on pp. 16, 83).

Daniel C. Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu and Ben Glocker (2019). 'Morpho-MNIST: quantitative assessment and diagnostics for representation learning'. *Journal of Machine Learning Research* 20, 178 (cit. on pp. 17, 103).

Daniel C. Castro, Ian Walker and Ben Glocker (2020). 'Causality matters in medical imaging'. *Nature Communications* 11, 3673. DOI: 10.1038/s41467-020-17478-w (cit. on pp. 15, 16, 19, 45).

Krzysztof Chalupka, Frederick Eberhardt and Pietro Perona (2017). 'Causal feature learning: an overview'. *Behaviormetrika* 44, pp. 137–164. DOI: 10.1007/s41237-016-0008-2 (cit. on p. 155).

Varun Chandola, Arindam Banerjee and Vipin Kumar (2009). 'Anomaly detection: a survey'. *ACM Computing Surveys* 41.3, pp. 1–58. DOI: 10.1145/1541880.1541882 (cit. on p. 94).

Olivier Chapelle, Bernhard Schölkopf and Alexander Zien, eds. (2006). *Semi-Supervised Learning*. Cambridge, MA: MIT Press (cit. on pp. 48, 50, 51, 90).

Ricky T. Q. Chen, Xuechen Li, Roger B. Grosse and David K. Duvenaud (2018a). 'Isolating sources of disentanglement in variational autoencoders'. In: *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pp. 2610–2620. arXiv: 1802.04942 (cit. on pp. 108, 117).

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt and David K. Duvenaud (2018b). 'Neural ordinary differential equations'. In: *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pp. 6572–6583 (cit. on p. 36).

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever and Pieter Abbeel (2016). 'InfoGAN: interpretable representation learning by information maximizing generative adversarial nets'. In: *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pp. 2172–2180 (cit. on pp. 114, 118, 135).

Silvia Chiappa and William S. Isaac (2019). 'A causal Bayesian networks viewpoint on fairness'. In: *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data (Privacy and Identity 2018)*. Vol. 547. IFIP AICT. Springer, Cham, pp. 3–20. DOI: 10.1007/978-3-030-16744-8_1 (cit. on pp. 60, 156).

Jae Young Choi, Wesley De Neve, Yong Man Ro and Konstantinos Plataniotis (2010). 'Automatic face annotation in personal photo collections using context-based unsupervised clustering and face information fusion'. *IEEE Transactions on Circuits and Systems for Video Technology* 20.10, pp. 1292–1309. DOI: 10.1109/TCSVT.2010.2058470 (cit. on p. 86).

Earl A. Coddington and Norman Levinson (1955). *Theory of Ordinary Differential Equations*. New York: McGraw-Hill (cit. on p. 36).

Gregory Cohen, Saeed Afshar, Jonathan Tapson and André van Schaik (2017). *EMNIST: an extension of MNIST to handwritten letters*. arXiv: 1702.05373. Dataset URL: https://www.nist.gov/itl/iad/image-group/emnist-dataset (cit. on p. 106).

Gary S. Collins and Karel G. M. Moons (2019). 'Reporting of artificial intelligence prediction models'. *The Lancet* 393.10181, pp. 1577–1579. DOI: 10.1016/S0140-6736(19)30037-6 (cit. on p. 60).

Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman and Karel G. M. Moons (2015). 'Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement'. *Annals of Internal Medicine* 162.1, pp. 55–63. DOI: 10.7326/M14-0697 (cit. on p. 60).

Robert J. Connor and James E. Mosimann (1969). 'Concepts of independence for proportions with a generalization of the Dirichlet distribution'. *Journal of the American Statistical Association* 64.325, pp. 194–206. DOI: 10.1080/01621459.1969.10500963 (cit. on p. 27).

M. Conroy, J. Sellors, M. Effingham, T. J. Littlejohns, C. Boultwood, L. Gillions, C. L. M. Sudlow, R. Collins and N. E. Allen (2019). 'The advantages of UK Biobank's open-access strategy for health research'. *Journal of Internal Medicine* 286.4, pp. 389–397. DOI: 10.1111/joim.12955 (cit. on p. 59).

David R. Cox (1955). 'Some statistical methods connected with series of events'. *Journal of the Royal Statistical Society: Series B (Methodological)* 17.2, pp. 129–164. JSTOR: 2983950 (cit. on p. 100).

Fabio Cozman and Ira Cohen (2006). 'Risks of semi-supervised learning: how unlabeled data can degrade performance of generative classifiers'. In: Chapelle, Schölkopf and Zien, 2006. Chap. 4, pp. 57–72 (cit. on p. 51).

Mario A. Cypko, Matthaeus Stoehr, Marcin Kozniewski, Marek J. Druzdzel, Andreas Dietz, Leonard Berliner and Heinz U. Lemke (2017). 'Validation workflow for a clinical Bayesian network model in multidisciplinary decision making in head and neck oncology treatment'. *International Journal of Computer Assisted Radiology and Surgery* 12.11, pp. 1959–1970. DOI: 10.1007/s11548-017-1531-7 (cit. on p. 47).

Andrew Dai and Amos J. Storkey (2015). 'The supervised hierarchical Dirichlet process'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.2, pp. 243–255. DOI: 10.1109/TPAMI.2014.2315802 (cit. on p. 178).

Rhian M. Daniel, Michael G. Kenward, Simon N. Cousens and Bianca L. De Stavola (2012). 'Using causal diagrams to guide analysis in missing data problems'. *Statistical Methods in Medical Research* 21.3, pp. 243–256. DOI: 10.1177/0962280210394469 (cit. on pp. 58, 154).

Povilas Daniušis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang and Bernhard Schölkopf (2010). 'Inferring deterministic causal relations'. In: *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010)*. AUAI Press, pp. 143–150 (cit. on p. 39).

Saloni Dash and Amit Sharma (2020). *Counterfactual generation and fairness evaluation using adversarially learned inference*. arXiv: 2009.08270 (cit. on pp. 132, 153).

Lee R. Dice (1945). 'Measures of the amount of ecologic association between species'. *Ecology* 26.3, pp. 297–302. DOI: 10.2307/1932409. JSTOR: 1932409 (cit. on p. 67).

Laurent Dinh, Jascha Sohl-Dickstein and Samy Bengio (2017). 'Density estimation using Real NVP'. In: *International Conference on Learning Representations (ICLR 2017)*. arXiv: 1605.0880 (cit. on p. 139).

Jeff Donahue, Philipp Krähenbühl and Trevor Darrell (2017). 'Adversarial feature learning'. In: *International Conference on Learning Representations (ICLR 2017)* (cit. on p. 132).

Qi Dou, **Daniel C. Castro**, Konstantinos Kamnitsas and Ben Glocker (2019). 'Domain generalization via model-agnostic learning of semantic features'. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* (Vancouver, Canada), pp. 6447–6458 (cit. on pp. 14, 17, 18).

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky and Aaron Courville (2017). 'Adversarially learned inference'. In: *International Conference on Learning Representations (ICLR 2017)* (cit. on p. 132).

Emilien Dupont (2018). 'Learning disentangled joint continuous and discrete representations'. In: *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pp. 708–718 (cit. on p. 108).

Conor Durkan, Artur Bekasov, Iain Murray and George Papamakarios (2019). 'Neural spline flows'. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 7511–7522 (cit. on pp. 35, 138).

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold and Aaron Roth (2015). 'The reusable holdout: preserving validity in adaptive data analysis'. *Science* 349.6248, pp. 636–638. DOI: 10.1126/science.aaa9375 (cit. on p. 48).

Cian Eastwood and Christopher K. I. Williams (2018). 'A framework for the quantitative evaluation of disentangled representations'. In: *International Conference on Learning Representations (ICLR 2018)* (cit. on pp. 105, 108, 117).

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau and Sebastian Thrun (2017). 'Dermatologist-level classification of skin cancer with deep neural networks'. *Nature* 542.7639, pp. 115–118. DOI: 10.1038/nature21056 (cit. on p. 45).

Warren J. Ewens (1972). 'The sampling theory of selectively neutral alleles'. *Theoretical Population Biology* 3.1, pp. 87–112. DOI: 10.1016/0040-5809(72)90035-4 (cit. on p. 25).

Thomas S. Ferguson (1973). 'A Bayesian analysis of some nonparametric problems'. *The Annals of Statistics* 1.2, pp. 209–230. JSTOR: 2958008 (cit. on pp. 23, 88).

Thomas S. Ferguson (1983). 'Bayesian density estimation by mixtures of normal distributions'. In: *Recent Advances in Statistics*, pp. 287–302. DOI: 10.1016/B978-0-12-589320-6.50018-6 (cit. on pp. 28, 64).

Adelino R. Ferreira da Silva (2007). 'A Dirichlet process mixture model for brain MRI tissue classification'. *Medical Image Analysis* 11.2, pp. 169–182. DOI: 10.1016/j.media.2006.12.002 (cit. on pp. 64, 67).

Adelino R. Ferreira da Silva (2009). 'Bayesian mixture models of variable dimension for image segmentation'. *Computer Methods and Programs in Biomedicine* 94.1, pp. 1–14. DOI: 10.1016/j.cmpb.2008.05.010 (cit. on pp. 64, 67).

Alejandro F. Frangi, Sotirios A. Tsaftaris and Jerry L. Prince (2018). 'Simulation and synthesis in medical imaging'. *IEEE Transactions on Medical Imaging* 37.3, pp. 673–679. DOI: 10.1109/TMI.2018.2800298 (cit. on p. 55).

Benoît Frénay and Michel Verleysen (2014). 'Classification in the presence of label noise: a survey'. *IEEE Transactions on Neural Networks and Learning Systems* 25.5, pp. 845–869. DOI: 10.1109/TNNLS.2013.2292894 (cit. on pp. 92, 178).

Andrew C. Gallagher and Tsuhan Chen (2009). 'Using context to recognize people in consumer images'. *IPSJ Transactions on Computer Vision and Applications* 1, pp. 115–126. DOI: 10.2197/ipsjtcva.1.115 (cit. on p. 86).

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March and Victor Lempitsky (2016). 'Domain-adversarial training of neural networks'. *Journal of Machine Learning Research* 17.59, pp. 1–35 (cit. on p. 55).

Leon A. Gatys, Alexander S. Ecker and Matthias Bethge (2016). 'Image style transfer using convolutional neural networks'. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 2414–2423. DOI: 10.1109/CVPR.2016.265 (cit. on p. 136).

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer W. Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford (2018). 'Datasheets for datasets'. In: *Proceedings of the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2018)*. arXiv: 1803.09010 (cit. on p. 60).

Markos Georgopoulos, Grigorios Chrysos, Maja Pantic and Yannis Panagakis (2020). 'Multilinear latent conditioning for generating unseen attribute combinations'. In: *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*. Vol. 119. PMLR. Forthcoming. PMLR. arXiv: 2009.04075 (cit. on p. 152).

Zoubin Ghahramani (2013). 'Bayesian non-parametrics and the probabilistic approach to modelling'. *Philosophical Transactions of the Royal Society A* 371.20110553. DOI: 10.1098/rsta.2011.0553 (cit. on pp. 22, 64).

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, Irene Y. Chen and Rajesh Ranganath (2019). 'Practical guidance on artificial intelligence for health-care data'. *The Lancet Digital Health* 1.4, e157–e159. DOI: 10.1016/S2589-7500(19)30084-6 (cit. on pp. 15, 59, 60, 154).

Ben Glocker, **Daniel C. Castro**, Robert Robinson and Ender Konukoglu (2019). 'Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects'. In: *Medical Imaging Meets NeurIPS Workshop* (Vancouver, Canada). arXiv: 1910.04597 (cit. on pp. 17, 18, 55, 81).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville and Yoshua Bengio (2014). 'Generative adversarial nets'. In: *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pp. 2672–2680 (cit. on pp. 108, 114, 135).

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz and Michèle Sebag (2018). 'Learning functional causal models with generative neural networks'. In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Ed. by Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Baró, Yağmur Güçlütürk, Umut Güçlü and Marcel van Gerven. Cham: Springer International Publishing, pp. 39–80. DOI: 10.1007/978-3-319-98131-4_3 (cit. on pp. 128, 130, 135).

Martin Grayson, Anja Thieme, Rita Marques, Daniela Massiceti, Ed Cutrell and Cecily Morrison (2020). 'A dynamic AI system for extending the capabilities of blind people'. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI EA '20. Honolulu, HI, USA: ACM, pp. 1–4. DOI: 10.1145/3334480.3383142 (cit. on p. 152).

Sander Greenland and Judea Pearl (2011). 'Causal diagrams'. In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Springer, Berlin, Heidelberg, pp. 208–216. DOI: 10.1007/978-3-642-04898-2_162 (cit. on pp. 40–42).

Sander Greenland, Judea Pearl and James M. Robins (1999). 'Causal diagrams for epidemiologic research'. *Epidemiology* 10.1, pp. 37–48. JSTOR: 3702180 (cit. on p. 128).

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf and Alexander J. Smola (2012). 'A kernel two-sample test'. *Journal of Machine Learning Research* 13.Mar, pp. 723–773 (cit. on pp. 114, 116).

Thomas Grote and Philipp Berens (2020). 'On the ethics of algorithmic decision-making in healthcare'. *Journal of Medical Ethics* 46.3, pp. 205–211. DOI: 10.1136/medethics-2019-105586 (cit. on p. 156).

Patrick J. Grother and Kayee K. Hanaoka (2016). *NIST Special Database 19*. Tech. rep. Gaithersburg, MD, USA: National Institute of Standards and Technology. DOI: 10.18434/T4H01C. Dataset URL: https://www.nist.gov/srd/nist-special-database-19 (cit. on p. 106).

Jason Hartford, Greg Lewis, Kevin Leyton-Brown and Matt Taddy (2017). 'Deep IV: a flexible approach for counterfactual prediction'. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. PMLR. PMLR, pp. 1414–1423 (cit. on p. 128).

Erion Hasanbelliu, Luis Sanchez Giraldo and José C. Principe (2011). 'A robust point matching algorithm for non-rigid registration using the Cauchy–Schwarz divergence'. In: *2011 IEEE International Workshop on Machine Learning for Signal Processing*. IEEE. DOI: 10.1109/MLSP.2011.6064593 (cit. on p. 72).

Pierre Hellier (2003). 'Consistent intensity correction of MR images'. In: *Proceedings of the 2003 International Conference on Image Processing (ICIP 2003)*. Vol. I. IEEE, pp. 1109–1112. DOI: 10.1109/ICIP.2003.1247161 (cit. on pp. 71, 72).

Miguel A. Hernán and Stephen R. Cole (2009). 'Causal diagrams and measurement bias'. *American Journal of Epidemiology* 170.8, pp. 959–962. DOI: 10.1093/aje/kwp293 (cit. on pp. 58, 155).

Miguel A. Hernán, Sonia Hernández-Díaz and James M. Robins (2004). 'A structural approach to selection bias'. *Epidemiology* 15.5, pp. 615–625. DOI: 10.1097/01.ede.0000135174.63482.43 (cit. on pp. 56, 57).

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler and Sepp Hochreiter (2017). 'GANs trained by a two time-scale update rule converge to a local Nash equilibrium'. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 6626–6637 (cit. on p. 108).

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed and Alexander Lerchner (2017). 'β-VAE: learning basic visual concepts with a constrained variational framework'. In: *International Conference on Learning Representations (ICLR 2017)* (cit. on pp. 108, 114, 135).

Matthew D. Hoffman, David M. Blei, Chong Wang and John Paisley (Jan. 2013). 'Stochastic variational inference'. *Journal of Machine Learning Research* 14.1, pp. 1303–1347 (cit. on pp. 29, 31).

Sharona Hoffman and Andy Podgurski (2013). 'The use and misuse of biomedical data: is bigger really better?' *American Journal of Law & Medicine* 39.4, pp. 497–538. DOI: 10.1177/009885881303900401 (cit. on pp. 60, 154).

Gary B. Huang, Manu Ramesh, Tamara Berg and Erik Learned-Miller (2007). *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Tech. rep. University of Massachusetts Amherst. Dataset URL: http://vis-www.cs.umass.edu/lfw/ (cit. on p. 93).

Lawrence Hubert and Phipps Arabie (1985). 'Comparing partitions'. *Journal of Classification* 2.1, pp. 193–218. DOI: 10.1007/BF01908075 (cit. on pp. 67, 95).

Yuankai Huo, Zhoubing Xu, Hyeonsoo Moon, Shunxing Bao, Albert Assad, Tamara K. Moyo, Michael R. Savona, Richard G. Abramson and Bennett A. Landman (2019). 'SynSeg-Net: synthetic segmentation without target modality ground truth'. *IEEE Transactions on Medical Imaging* 38.4, pp. 1016–1025. DOI: 10.1109/TMI.2018.2876633 (cit. on p. 55).

Maximilian Ilse, Jakub M. Tomczak and Patrick Forré (2020). *Designing data augmentation for simulating interventions*. arXiv: 2005.01856 (cit. on pp. 128, 155).

Sergey Ioffe (2006). 'Probabilistic linear discriminant analysis'. In: *Computer Vision – ECCV 2006*. Vol. 3954. LNCS, pp. 531–542. DOI: 10.1007/11744085_41 (cit. on p. 84).

Hemant Ishwaran and Mahmoud Zarepour (2000). 'Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models'. *Biometrika* 87.2, pp. 371–390. DOI: 10.1093/biomet/87.2.371 (cit. on p. 27).

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou and Alexei A. Efros (2017). 'Image-to-image translation with conditional adversarial networks'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5967–5976. DOI: 10.1109/CVPR.2017.632 (cit. on p. 136).

Rabia Jafri and Hamid R. Arabnia (2009). 'A survey of face recognition techniques'. *Journal of Information Processing Systems* 5.2, pp. 41–68. DOI: 10.3745/JIPS.2009.5.2.041 (cit. on pp. 84, 86).

Eric Jang, Shixiang Gu and Ben Poole (2017). 'Categorical reparameterization with Gumbel-softmax'. In: *International Conference on Learning Representations (ICLR 2017)* (cit. on pp. 131, 135).

Saad Jbabdi, Mark W. Woolrich and Timothy E. J. Behrens (2009). 'Multiple-subjects connectivity-based parcellation using hierarchical Dirichlet process mixture models'. *NeuroImage* 44.2, pp. 373–384. DOI: 10.1016/j.neuroimage.2008.08.044 (cit. on p. 64).

Bing Jian and Baba C. Vemuri (2011). 'Robust point set registration using Gaussian mixture models'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.8, pp. 1633–1645. DOI: 10.1109/TPAMI.2010.223 (cit. on p. 72).

Matthew J. Johnson, David K. Duvenaud, Alex Wiltschko, Ryan P. Adams and Sandeep R. Datta (2016). 'Composing graphical models with neural networks for structured representations and fast inference'. In: *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pp. 2946–2954 (cit. on pp. 32, 135).

Indra Joshi and Jessica Morley, eds. (Oct. 2019). *Artificial Intelligence: How to get it right. Putting policy into practice for safe data-driven innovation in health and care*. London, UK: NHSX. URL: https://www.nhsx.nhs.uk/media/documents/NHSX_AI_report.pdf (cit. on p. 59).

Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert et al. (2017a). 'Unsupervised domain adaptation in brain lesion segmentation with adversarial networks'. In: *Information Processing in Medical Imaging*. Vol. 10265. LNCS. Springer, Cham, pp. 597–609. DOI: 10.1007/978-3-319-59050-9_47 (cit. on p. 55).

Konstantinos Kamnitsas, **Daniel C. Castro**, Loïc Le Folgoc, Ian Walker, Ryutaro Tanno, Daniel Rueckert, Ben Glocker, Antonio Criminisi and Aditya Nori (2018). 'Semi-supervised learning via compact latent space clustering'. In: *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)* (Stockholm, Sweden). Vol. 80. PMLR. PMLR, pp. 2464–2473 (cit. on pp. 17, 18).

Konstantinos Kamnitsas, Christian Ledig, Virginia F. J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert and Ben Glocker (2017b). 'Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation'. *Medical Image Analysis* 36, pp. 61–78. DOI: 10.1016/j.media.2016.10.004 (cit. on pp. 64, 69).

Divyansh Kaushik, Eduard Hovy and Zachary C. Lipton (2020). 'Learning the difference that makes a difference with counterfactually-augmented data'. In: *International Conference on Learning Representations*. arXiv: 1909.12434 (cit. on p. 128).

Alex Kendall and Yarin Gal (2017). 'What uncertainties do we need in Bayesian deep learning for computer vision?' In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 5574–5584 (cit. on p. 122).

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing and Bernhard Schölkopf (2017). 'Avoiding discrimination through causal reasoning'. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 656–666 (cit. on pp. 60, 156).

Hyunjik Kim and Andriy Mnih (2018). 'Disentangling by factorising'. In: *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*. Vol. 80. PMLR. PMLR, pp. 2649–2658 (cit. on p. 119).

Seyoung Kim and Padhraic Smyth (2006). 'Hierarchical Dirichlet processes with random effects'. In: *Advances in Neural Information Processing Systems 19 (NIPS 2006)*. MIT Press, pp. 697–704 (cit. on p. 64).

Diederik P. Kingma and Jimmy Ba (2014). 'Adam: a method for stochastic optimization'. In: *International Conference on Learning Representations (ICLR 2015)*. arXiv: 1412.6980 (cit. on p. 139).

Diederik P. Kingma and Max Welling (2014a). 'Auto-encoding variational Bayes'. In: *International Conference on Learning Representations (ICLR 2014)*. arXiv: 1312.6114 (cit. on pp. 33, 107, 131, 135).

Diederik P. Kingma and Max Welling (2014b). 'Efficient gradient-based inference through transformations between Bayes nets and neural nets'. In: *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*. Vol. 32. PMLR 2. PMLR, pp. 1782–1790 (cit. on p. 33).

Ivan Kobyzev, Simon Prince and Marcus Brubaker (2020). 'Normalizing flows: an introduction and review of current methods'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. To appear. DOI: 10.1109/tpami.2020.2992934 (cit. on p. 34).

Murat Kocaoglu, Christopher Snyder, Alexandros G. Dimakis and Sriram Vishwanath (2018). 'CausalGAN: learning causal implicit generative models with adversarial training'. In: *International Conference on Learning Representations* (cit. on pp. 128, 130, 135, 136).

Tejas D. Kulkarni, William F. Whitney, Pushmeet Kohli and Josh Tenenbaum (2015). 'Deep convolutional inverse graphics network'. In: *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pp. 2539–2547 (cit. on p. 135).

Abhishek Kumar, Prasanna Sattigeri and Avinash Balakrishnan (2018). 'Variational inference of disentangled latent concepts from unlabeled observations'. In: *International Conference on Learning Representations (ICLR 2018)* (cit. on pp. 108, 117).

Matt Kusner, Joshua Loftus, Chris Russell and Ricardo Silva (2017). 'Counterfactual fairness'. In: *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 4066–4076 (cit. on pp. 60, 128, 156).

Brenden M. Lake, Ruslan R. Salakhutdinov and Joshua B. Tenenbaum (2015). 'Human-level concept learning through probabilistic program induction'. *Science* 350.6266, pp. 1332–1338. DOI: 10.1126/science.aab3050. Dataset URL: https://github.com/brendenlake/omniglot (cit. on p. 124).

Curtis P. Langlotz, Bibb Allen, Bradley J. Erickson, Jayashree Kalpathy-Cramer, Keith Bigelow, Tessa S. Cook, Adam E. Flanders, Matthew P. Lungren, David S. Mendelson, Jeffrey D. Rudie et al. (2019). 'A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy Workshop'. *Radiology* 291.3, pp. 781–791. DOI: 10.1148/radiol.2019190613 (cit. on pp. 59, 60).

Hugo Larochelle, Dumitru Erhan, Aaron Courville, James Bergstra and Yoshua Bengio (2007). 'An empirical evaluation of deep architectures on problems with many factors of variation'. In: *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*, pp. 473–480. DOI: 10.1145/1273496.1273556. Dataset URL: https://www.iro.umontreal.ca/~lisa/twiki/bin/view.cgi/Public/MnistVariations (cit. on p. 107).

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle and Ole Winther (2016). 'Autoencoding beyond pixels using a learned similarity metric'. In: *Proceedings of The 33rd International Conference on Machine Learning (ICML 2016)*. Vol. 48. PMLR. PMLR, pp. 1558–1566 (cit. on p. 134).

Timothy L. Lash, Matthew P. Fox, Richard F. MacLehose, George Maldonado, Lawrence C. McCandless and Sander Greenland (2014). 'Good practices for quantitative bias analysis'. *International Journal of Epidemiology* 43.6, pp. 1969–1985. DOI: 10.1093/ije/dyu149 (cit. on p. 58).

Loïc Le Folgoc, **Daniel C. Castro**, Jeremy Tan, Bishesh Khanal, Konstantinos Kamnitsas, Ian Walker, Amir Alansary and Ben Glocker (2019). 'Controlling meshes via curvature: spin transformations for pose-invariant shape processing'. In: *Information Processing in Medical Imaging* (Hong Kong). Vol. 11492. LNCS. Springer, Cham, pp. 221–234. DOI: 10.1007/978-3-030-20351-1_17. arXiv: 1903.02429 (cit. on p. 18).

Nam Le, Hervé Bredin, Gabriel Sargent, Miquel India, Paula Lopez-Otero, Claude Barras, Camille Guinaudeau, Guillaume Gravier, Gabriel Barbosa da Fonseca, Izabela Lyon Freire et al. (2017). 'Towards large scale multimedia indexing: a case study on person discovery in broadcast news'. In: *Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing (CBMI 2017)*. Florence, Italy: ACM, 18:1–18:6. DOI: 10.1145/3095713.3095732 (cit. on p. 86).

Erik Learned-Miller and Gary B. Huang (2016). 'Labeled Faces in the Wild: a survey'. In: *Advances in Face Detection and Facial Image Analysis*. Ed. by Michal Kawulok, M. Emre Celebi and Bogdan Smolka. Springer, pp. 189–248. DOI: 10.1007/978-3-319-25958-1 (cit. on p. 85).

Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, R. E. Howard, Wayne E. Hubbard and Lawrence D. Jackel (1990). 'Handwritten digit recognition with a back-propagation network'. In: *Advances in Neural Information Processing Systems 2 (NIPS 1989)*, pp. 396–404. Dataset URL: https://web.stanford.edu/~hastie/ElemStatLearn/data.html (cit. on p. 124).

Yann LeCun, Léon Bottou, Yoshua Bengio and Patrick Haffner (1998). 'Gradient-based learning applied to document recognition'. *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791. Dataset URL: http://yann.lecun.com/exdb/mnist/ (cit. on pp. 101, 104, 106, 107, 111, 112, 121, 122, 136, 137).

Christian Ledig, Rolf A. Heckemann, Alexander Hammers, Juan Carlos Lopez, Virginia F.J. Newcombe, Antonios Makropoulos, Jyrki Lötjönen, David K. Menon and Daniel Rueckert (2015). 'Robust whole-brain segmentation: application to traumatic brain injury'. *Medical Image Analysis* 21.1, pp. 40–58. DOI: 10.1016/j.media.2014.12.003 (cit. on p. 64).

Christian Ledig, Robin Wolz, Paul Aljabar, Jyrki Lötjönen, Rolf A. Heckemann, Alexander Hammers and Daniel Rueckert (2012). 'Multi-class brain segmentation using atlas propagation and EM-based refinement'. In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 896–899. DOI: 10.1109/ISBI.2012.6235693 (cit. on p. 64).

Chongxuan Li, Max Welling, Jun Zhu and Bo Zhang (2018). 'Graphical generative adversarial networks'. In: *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pp. 6069–6080 (cit. on p. 135).

Peng Li, Yun Fu, Umar Mohammed, James H. Elder and Simon J. D. Prince (2012). 'Probabilistic models for inference about identity'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.1, pp. 144–157. DOI: 10.1109/TPAMI.2011.104 (cit. on p. 84).

Wu Lin, Nicolas Hubacher and Mohammad Emtiyaz Khan (2018). 'Variational message passing with structured inference networks'. In: *International Conference on Learning Representations (ICLR 2018)* (cit. on pp. 32, 135).

Zachary C. Lipton, Yu-Xiang Wang and Alexander J. Smola (2018). 'Detecting and correcting for label shift with black box predictors'. In: *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*. Vol. 80. PMLR. PMLR, pp. 3122–3130 (cit. on p. 55).

Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J. Calvert, Alastair K. Denniston and The SPIRIT-AI and CONSORT-AI Working Group (2020). 'Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension'. *Nature Medicine* 26, pp. 1364–1374. DOI: 10.1038/s41591-020-1034-x (cit. on pp. 59, 60).

James R. Lloyd and Zoubin Ghahramani (2015). 'Statistical model criticism using kernel two sample tests'. In: *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pp. 829–837 (cit. on p. 116).

Gaëlle Loosli, Stéphane Canu and Léon Bottou (2007). 'Training invariant support vector machines using selective sampling'. In: *Large Scale Kernel Machines*. Ed. by Léon Bottou, Olivier Chapelle, Dennis DeCoste and Jason Weston. Cambridge, MA, USA: MIT Press, pp. 301–320. Dataset URL: http://leon.bottou.org/projects/infimnist (cit. on p. 107).

Mohammad Lotfollahi, Mohsen Naghipourfar, Fabian J. Theis and F. Alexander Wolf (2019). *Conditional out-of-sample generation for unpaired data using trVAE*. arXiv: 1910.01791 (cit. on p. 152).

You Lu and Bert Huang (2020). 'Structured output learning with conditional generative flows'. In: *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)*. Vol. 34. 4. AAAI, pp. 5005–5012. DOI: 10.1609/aaai.v34i04.5940 (cit. on pp. 35, 131).

Peter J. F. Lucas, Linda C. van der Gaag and Ameen Abu-Hanna (2004). 'Bayesian networks in biomedicine and health-care'. *Artificial Intelligence in Medicine* 30.3, pp. 201–214. DOI: 10.1016/j.artmed.2003.11.001 (cit. on p. 47).

Chris J. Maddison, Andriy Mnih and Yee Whye Teh (2017). 'The Concrete distribution: a continuous relaxation of discrete random variables'. In: *International Conference on Learning Representations (ICLR 2017)* (cit. on pp. 131, 135).

Chris J. Maddison and Daniel Tarlow (Jan. 2017). *Gumbel Machinery*. URL: https://cmaddis.github.io/gumbel-machinery (visited on 14/05/2020) (cit. on pp. 134, 135).

Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass et al. (2018). 'Why rankings of biomedical image analysis competitions should be interpreted with care'. *Nature Communications* 9.5217. DOI: 10.1038/s41467-018-07619-7 (cit. on p. 49).

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow and Brendan Frey (2016). 'Adversarial autoencoders'. In: *International Conference on Learning Representations Workshop (ICLRW 2016)*. arXiv: 1511.05644 (cit. on p. 108).

Kanti V. Mardia and Peter E. Jupp (1999). 'Distributions on spheres'. In: *Directional Statistics*. John Wiley & Sons, Inc. Chap. 9, pp. 159–192. DOI: 10.1002/9780470316979.ch9 (cit. on p. 100).

Álvaro Parafita Martínez and Jordi Vitrià Marca (2019). 'Explaining visual models by causal attribution'. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, pp. 4167–4175. DOI: 10.1109/ICCVW.2019.00512 (cit. on pp. 128, 136).

Loic Matthey, Irina Higgins, Demis Hassabis and Alexander Lerchner (2017). *dSprites: Disentanglement testing Sprites dataset*. URL: https://github.com/deepmind/dsprites-dataset/ (visited on 08/05/2018) (cit. on pp. 107, 124).

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest et al. (2015). 'The multimodal brain tumor image segmentation benchmark (BRATS)'. *IEEE Transactions on Medical Imaging* 34.10, pp. 1993–2024. DOI: 10.1109/TMI.2014.2377694 (cit. on p. 45).

Karla L. Miller, Fidel Alfaro-Almagro, Neal K. Bangerter, David L. Thomas, Essa Yacoub, Junqian Xu, Andreas J. Bartsch, Saad Jbabdi, Stamatios N. Sotiropoulos, Jesper L. R. Andersson et al. (2016). 'Multimodal population brain imaging in the UK Biobank prospective epidemiological study'. *Nature Neuroscience* 19.11, pp. 1523–1536. DOI: 10.1038/nn.4393 (cit. on p. 59).

Mehdi Mirza and Simon Osindero (2014). *Conditional generative adversarial nets*. arXiv: 1411.1784 (cit. on pp. 132, 135).

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji and Timnit Gebru (2019). 'Model cards for model reporting'. In: *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency (FAT* 2019)*. ACM, pp. 220–229. DOI: 10.1145/3287560.3287596 (cit. on p. 60).

Shakir Mohamed, Mihaela Rosca, Michael Figurnov and Andriy Mnih (2020). 'Monte Carlo gradient estimation in machine learning'. *Journal of Machine Learning Research* 21.132 (cit. on p. 32).

Karthika Mohan, Judea Pearl and Jin Tian (2013). 'Graphical models for inference with missing data'. In: *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pp. 1277–1285 (cit. on pp. 58, 154).

Miguel Monteiro, Loïc Le Folgoc, **Daniel C. Castro**, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk and Ben Glocker (2020). 'Stochastic segmentation networks: modelling spatially correlated aleatoric uncertainty'. In: *Advances in Neural Information Processing Systems (NeurIPS 2020)*. Forthcoming. arXiv: 2006.06015 (cit. on p. 18).

Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin and Huan Liu (2020). 'Causal interpretability for machine learning – Problems, methods and evaluation'. *ACM SIGKDD Explorations Newsletter* 22.1, pp. 18–33. DOI: 10.1145/3400051 (cit. on p. 156).

Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla and Francisco Herrera (2012). 'A unifying view on dataset shift in classification'. *Pattern Recognition* 45.1, pp. 521–530. DOI: 10.1016/j.patcog.2011.06.019 (cit. on pp. 55, 57).

Radford M. Neal (2000). 'Markov chain sampling methods for Dirichlet process mixture models'. *Journal of Computational and Graphical Statistics* 9.2, pp. 249–265. DOI: 10.1080/10618600.2000.10474879 (cit. on p. 99).

Sebastian Nowozin, Tom Ellis, Cecily Morrison and **Daniel C. Castro** (2019). 'Image Processing for Person Recognition'. U.S. patent 10621416. Microsoft Technology Licensing, LLC (cit. on pp. 16, 83).

László G. Nyúl and Jayaram K. Udupa (1999). 'On standardizing the MR image intensity scale'. *Magnetic Resonance in Medicine* 42.6, pp. 1072–1081 (cit. on p. 71).

László G. Nyúl, Jayaram K. Udupa and Xuan Zhang (2000). 'New variants of a method of MRI scale standardization'. *IEEE Transactions on Medical Imaging* 19.2, pp. 143–150. DOI: 10.1109/42.836373 (cit. on pp. 71, 78, 81).

Michael Oberst and David Sontag (2019). 'Counterfactual off-policy evaluation with Gumbel-Max structural causal models'. In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. PMLR. PMLR, pp. 4881–4890 (cit. on p. 134).

Avital Oliver, Augustus Odena, Colin A. Raffel, Ekin D. Cubuk and Ian Goodfellow (2018). 'Realistic evaluation of deep semi-supervised learning algorithms'. In: *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pp. 3235–3246 (cit. on p. 51).

George Papamakarios, Eric Nalisnick, Danilo J. Rezende, Shakir Mohamed and Balaji Lakshminarayanan (2019). *Normalizing flows for probabilistic modeling and inference*. arXiv: 1912.02762 (cit. on p. 34).

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla and Bernhard Schölkopf (2018). 'Learning independent causal mechanisms'. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. PMLR. PMLR, pp. 4036–4044 (cit. on p. 128).

Ravi B. Parikh, Ziad Obermeyer and Amol S. Navathe (2019). 'Regulation of predictive analytics in medicine'. *Science* 363.6429, pp. 810–812. DOI: 10.1126/science.aaw0029 (cit. on p. 59).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga et al. (2019). 'PyTorch: an imperative style, high-performance deep learning library'. In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 8024–8035 (cit. on p. 137).

Ashis Pati and Alexander Lerch (2020). 'Attribute-based regularization of latent spaces for variational auto-encoders'. *Neural Computing and Applications*. DOI: 10.1007/s00521-020-05270-2 (cit. on pp. 135, 152).

Nick Pawlowski*, **Daniel C. Castro*** and Ben Glocker (2020). 'Deep structural causal models for tractable counterfactual inference'. In: *Advances in Neural Information Processing Systems (NeurIPS 2020)*. Forthcoming, *equal contribution. arXiv: 2006.06485 (cit. on pp. 17, 127, 132).

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani and Thomas Vetter (2009). 'A 3D face model for pose and illumination invariant face recognition'. In: *Proceedings of the Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS 2009)*. IEEE, pp. 296–301. DOI: 10.1109/AVSS.2009.58. Dataset URL: https://faces.dmi.unibas.ch/bfm/index.php?nav=1-1-1&id=scans (cit. on p. 107).

Judea Pearl (2009). *Causality: Models, Reasoning, and Inference*. 2nd. Cambridge University Press (cit. on pp. 38–41, 57, 128, 130, 153).

Judea Pearl (2019). 'The seven tools of causal inference, with reflections on machine learning'. *Communications of the ACM* 62.3, pp. 54–60. DOI: 10.1145/3241036 (cit. on pp. 37, 129).

Judea Pearl and Elias Bareinboim (2014). 'External validity: from do-calculus to transportability across populations'. *Statistical Science* 29.4, pp. 579–595. DOI: 10.1214/14-STS486 (cit. on pp. 42, 54).

Judea Pearl and Dana Mackenzie (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books (cit. on p. 37).

Serafeim Perdikis, Robert Leeb, Ricardo Chavarriaga and José del R. Millán (2015). *Context-aware learning for finite mixture models*. arXiv: 1507.08272 (cit. on p. 88).

Jonas Peters, Dominik Janzing and Bernhard Schölkopf (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: MIT Press (cit. on pp. 39, 41, 128–130, 153).

Ruth Peters (2006). 'Ageing and the brain'. *Postgraduate Medical Journal* 82.964, pp. 84–88. DOI: 10.1136/pgmj.2005.036665 (cit. on pp. 144, 147).

Jim Pitman (2006). *Combinatorial Stochastic Processes. Ecole d'Eté de Probabilités de Saint-Flour XXXII – 2002*. Ed. by Jean Picard. Vol. 1875. Lecture Notes in Mathematics. Berlin: Springer, Berlin, Heidelberg. DOI: 10.1007/b11601500 (cit. on pp. 27, 89).

Jim Pitman and Marc Yor (1997). 'The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator'. *The Annals of Probability* 25.2, pp. 855–900. JSTOR: 2959614 (cit. on pp. 27, 91).

Vinay Uday Prabhu (2019). *Kannada-MNIST: a new handwritten digits dataset for the Kannada language*. arXiv: 1908.01242 (cit. on p. 152).

Luciano M. Prevedello, Safwan S. Halabi, George Shih, Carol C. Wu, Marc D. Kohli, Falgun H. Chokshi, Bradley J. Erickson, Jayashree Kalpathy-Cramer, Katherine P. Andriole and Adam E. Flanders (2019). 'Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions'. *Radiology: Artificial Intelligence* 1.1, e180031. DOI: 10.1148/ryai.2019180031 (cit. on pp. 59, 60).

Simon J.D. Prince and James H. Elder (2007). 'Probabilistic linear discriminant analysis for inferences about identity'. In: *Proceedings of the 11th IEEE International Conference on Computer Vision (ICCV 2007)*. IEEE. DOI: 10.1109/ICCV.2007.4409052 (cit. on p. 84).

José C. Principe, Dongxin Xu and John W. Fisher III (2000). 'Information-theoretic learning'. In: *Unsupervised Adaptive Filtering*. Vol. 1. John Wiley & Sons, Inc. Chap. 7 (cit. on p. 73).

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer and Neil D. Lawrence, eds. (2009). *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press (cit. on pp. 53, 54).

William M. Rand (1971). 'Objective criteria for the evaluation of clustering methods'. *Journal of the American Statistical Association* 66.336, pp. 846–850. DOI: 10.1080/01621459.1971.10482356 (cit. on pp. 67, 95).

Carl E. Rasmussen (2000). 'The infinite Gaussian mixture model'. In: *Advances in Neural Information Processing Systems 12 (NIPS 1999)*, pp. 554–560 (cit. on p. 28).

Hans Reichenbach (1956). *The Direction of Time*. Berkeley, CA: University of California Press (cit. on p. 39).

Danilo J. Rezende and Shakir Mohamed (2015). 'Variational inference with normalizing flows'. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*. Vol. 37. PMLR. PMLR, pp. 1530–1538 (cit. on p. 135).

Danilo J. Rezende, Shakir Mohamed and Daan Wierstra (2014). 'Stochastic backpropagation and approximate inference in deep generative models'. In: *Proceedings of the 31st International Conference on Machine Learning (ICML 2014)*. Vol. 32. PMLR 2. PMLR, pp. 1278–1286 (cit. on pp. 32, 33, 131).

Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K. Denniston, Melanie J. Calvert, The SPIRIT-AI and CONSORT-AI Working Group, SPIRIT-AI and CONSORT-AI Steering Group and SPIRIT-AI and CONSORT-AI Consensus Group (2020). 'Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension'. *Nature Medicine* 26, pp. 1351–1363. DOI: 10.1038/s41591-020-1037-7 (cit. on pp. 59, 60).

Robert Robinson, Qi Dou, **Daniel C. Castro**, Konstantinos Kamnitsas, Marius de Groot, Ronald M. Summers, Daniel Rueckert and Ben Glocker (2020). 'Image-level harmonization of multi-site data using image-and-spatial transformer networks'. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2020*. Forthcoming. arXiv: 2006.16741 (cit. on pp. 17, 18, 153).

Abel Rodríguez, David B. Dunson and Alan E. Gelfand (2008). 'The nested Dirichlet process'. *Journal of the American Statistical Association* 103.483, pp. 1131–1154. DOI: 10.1198/016214508000000553 (cit. on p. 99).

Andrew Rosenberg and Julia Hirschberg (2007). 'V-measure: a conditional entropy-based external cluster evaluation measure'. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 410–420 (cit. on p. 67).

Arunabha S. Roy, Ajay Gopinath and Anand Rangarajan (2007). 'Deformable density matching for 3D non-rigid registration of shapes'. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007*. Vol. 4791. LNCS. Springer, Berlin, Heidelberg, pp. 942–949. DOI: 10.1007/978-3-540-75757-3_114 (cit. on p. 72).

Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup and Bernhard Schölkopf (2017). 'Causal consistency of structural equation models'. In: *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI 2017)*. AUAI Press (cit. on p. 155).

Donald B. Rubin (1976). 'Inference and missing data'. *Biometrika* 63.3, 581. DOI: 10.2307/2335739. JSTOR: 2335739 (cit. on p. 58).

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford and Xi Chen (2016). 'Improved techniques for training GANs'. In: *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pp. 2234–2242 (cit. on p. 108).

Suchi Saria and Adarsh Subbaswamy (2019). *Tutorial: safe and reliable machine learning*. Tutorial presented at the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT* 2019). arXiv: 1904.07204 (cit. on p. 60).

Jo Schlemper, **Daniel C. Castro**, Wenjia Bai, Chen Qin, Ozan Oktay, Jinming Duan, Anthony N. Price, Joseph V. Hajnal and Daniel Rueckert (2018a). 'Bayesian deep learning for accelerated MR image reconstruction'. In: *International Workshop on Machine Learning for Medical Image Reconstruction* (Granada, Spain). Vol. 11074. LNCS. Springer, Cham, pp. 64–71. DOI: 10.1007/978-3-030-00129-2_8 (cit. on p. 18).

Jo Schlemper, Ozan Oktay, Wenjia Bai, **Daniel C. Castro**, Jinming Duan, Chen Qin, Joseph V. Hajnal and Daniel Rueckert (2018b). 'Cardiac MR segmentation from undersampled k-space using deep latent representation learning'. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018* (Granada, Spain). Vol. 11070. LNCS. Springer, Cham, pp. 259–267. DOI: 10.1007/978-3-030-00928-1_30 (cit. on p. 18).

Bernhard Schölkopf (2019). *Causality for machine learning*. arXiv: 1911.10500 (cit. on p. 128).

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang and Joris Mooij (2012). 'On causal and anticausal learning'. In: *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pp. 459–466 (cit. on pp. 48, 49).

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang and Joris Mooij (2013). 'Semi-supervised learning in causal and anticausal settings'. In: *Empirical Inference*. Ed. by Bernhard Schölkopf, Zhiyuan Luo and Vladimir Vovk. Springer, Berlin, Heidelberg. Chap. 13, pp. 129–141. DOI: 10.1007/978-3-642-41136-6_13 (cit. on pp. 48, 51).

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola and Robert C. Williamson (2001). 'Estimating the support of a high-dimensional distribution'. *Neural Computation* 13.7, pp. 1443–1471. DOI: 10.1162/089976601750264965 (cit. on p. 94).

Florian Schroff, Dmitry Kalenichenko and James Philbin (2015). 'FaceNet: a unified embedding for face recognition and clustering'. In: *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*. IEEE, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682 (cit. on pp. 85, 89).

David W. Scott (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. New York, NY, USA: John Wiley & Sons, Inc. (cit. on p. 116).

Jayaram Sethuraman (1994). 'A constructive definition of Dirichlet priors'. *Statistica Sinica* 4.2, pp. 639–650. JSTOR: 24305538 (cit. on pp. 27, 89).

Mohak Shah, Yiming Xiao, Nagesh Subbanna, Simon Francis, Douglas L. Arnold, D. Louis Collins and Tal Arbel (2011). 'Evaluating intensity normalization on MRIs of human brain with multiple sclerosis'. *Medical Image Analysis* 15.2, pp. 267–282. DOI: 10.1016/j.media.2010.12.003 (cit. on p. 71).

Nigam H. Shah, Arnold Milstein and Steven C. Bagley (2019). 'Making machine learning models clinically useful'. *Journal of the American Medical Association* 322.14, pp. 1351–1352. DOI: 10.1001/jama.2019.10306 (cit. on pp. 59, 60).

Nilay D. Shah, Ewout W. Steyerberg and David M. Kent (2018). 'Big data and predictive analytics: recalibrating expectations'. *Journal of the American Medical Association* 320.1, pp. 27–28. DOI: 10.1001/jama.2018.5602 (cit. on pp. 59, 60, 154).

Eyal Shahar (2009). 'Causal diagrams for encoding and evaluation of information bias'. *Journal of Evaluation in Clinical Practice* 15.3, pp. 436–440. DOI: 10.1111/j.1365-2753.2008.01031.x (cit. on pp. 58, 155).

Chengzhi Shi, Ben Glocker and **Daniel C. Castro** (2019). 'PVAE: learning disentangled representations with intrinsic dimension via approximated L_0 regularization'. In: *NeurIPS 2019 Disentanglement Challenge*. eprint: https://openreview.net/forum?id=HJg8stY2oB (cit. on p. 18).

Aarti Singh, Robert Nowak and Xiaojin Zhu (2008). 'Unlabeled data: now it helps, now it doesn't'. In: *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, pp. 1513–1520 (cit. on p. 51).

Sumedha Singla, Brian Pollack, Junxiang Chen and Kayhan Batmanghelich (2020). 'Explanation by progressive exaggeration'. In: *International Conference on Learning Representations (ICLR 2020)* (cit. on pp. 128, 136, 150).

Lewis Smith and Yarin Gal (2018). 'Understanding measures of uncertainty for adversarial example detection'. In: *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI 2018)*. AUAI Press, pp. 560–569 (cit. on p. 122).

Kihyuk Sohn, Honglak Lee and Xinchen Yan (2015). 'Learning structured output representation using deep conditional generative models'. In: *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pp. 3483–3491 (cit. on p. 135).

Amos J. Storkey (2009). 'When training and test sets are different: characterising learning transfer'. In: Quiñonero-Candela, Sugiyama, Schwaighofer and Lawrence, 2009. Chap. 1, pp. 3–28 (cit. on pp. 14, 54).

Julian Straub, Jason Chang, Oren Freifeld and John W. Fisher III (2015). 'A Dirichlet process mixture model for spherical data'. In: *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS 2015)*. Vol. 38. PMLR. PMLR, pp. 930–938 (cit. on p. 100).

Adarsh Subbaswamy, Peter Schulam and Suchi Saria (2019). 'Preventing failures due to dataset shift: learning predictive models that transport'. In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*. Vol. 89. PMLR. PMLR, pp. 3118–3127 (cit. on pp. 14, 60, 128).

Erik B. Sudderth and Michael I. Jordan (2009). 'Shared segmentation of natural scenes using dependent Pitman–Yor processes'. In: *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, pp. 1585–1592 (cit. on p. 64).

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray et al. (2015). 'UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age'. *PLoS Medicine* 12.3. DOI: 10.1371/journal.pmed.1001779 (cit. on p. 144).

Dougal J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alexander J. Smola and Arthur Gretton (2017). 'Generative models and model criticism via optimized maximum mean discrepancy'. In: *International Conference on Learning Representations (ICLR 2017)*. arXiv: 1611.04488 (cit. on p. 116).

Jeffrey W. Swanson and Jennifer K. Ibrahim (2013). 'Picturing public health law research: the value of causal diagrams'. In: *Public Health Law Research: Theory and Methods*. Ed. by C. Alexander Wagenaar and Scott Burris. Jossey-Bass. Chap. 10, pp. 217–236. SSRN: 2027558 (cit. on pp. 42, 43).

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens and Zbigniew Wojna (2016). 'Rethinking the Inception architecture for computer vision'. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*. IEEE, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308 (cit. on p. 108).

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow and Rob Fergus (2014). 'Intriguing properties of neural networks'. In: *International Conference on Learning Representations (ICLR 2014)*. arXiv: 1312.6199 (cit. on p. 128).

Esteban G. Tabak and Cristina V. Turner (2013). 'A family of nonparametric density estimation algorithms'. *Communications on Pure and Applied Mathematics* 66.2, pp. 145–164. DOI: 10.1002/cpa.21423 (cit. on p. 35).

Esteban G. Tabak and Eric Vanden-Eijnden (2010). 'Density estimation by dual ascent of the log-likelihood'. *Communications in Mathematical Sciences* 8.1, pp. 217–233 (cit. on p. 34).

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato and Lior Wolf (2014). 'DeepFace: closing the gap to human-level performance in face verification'. In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*. IEEE, pp. 1701–1708. DOI: 10.1109/CVPR.2014.220 (cit. on pp. 85, 90).

Makarand Tapaswi, Martin Bäuml and Rainer Stiefelhagen (2012). '"Knock! Knock! Who is it?" Probabilistic person identification in TV-series'. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, pp. 2658–2665. DOI: 10.1109/CVPR.2012.6247986 (cit. on p. 86).

Yee Whye Teh (2010). 'Dirichlet process'. In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Springer US, pp. 280–287. DOI: 10.1007/978-0-387-30164-8_219 (cit. on pp. 23, 26, 88, 91).

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal and David M. Blei (2006). 'Hierarchical Dirichlet processes'. *Journal of the American Statistical Association* 101.476, pp. 1566–1581. DOI: 10.1198/016214506000000302 (cit. on pp. 88, 89, 99, 178, 179).

Loo-Nin Teow and Kia-Fock Loe (2002). 'Robust vision-based features and classification schemes for off-line handwritten digit recognition'. *Pattern Recognition* 35.11, pp. 2355–2364. DOI: 10.1016/S0031-3203(01)00228-X (cit. on p. 111).

Lucas Theis, Aäron van den Oord and Matthias Bethge (2016). 'A note on the evaluation of generative models'. In: *International Conference on Learning Representations (ICLR 2016)*. arXiv: 1511.01844 (cit. on pp. 108, 114).

Bertrand Thirion, Alan Tucholka, Merlin Keller, Philippe Pinel, Alexis Roche, Jean-François Mangin and Jean-Baptiste Poline (2007). 'High level group analysis of FMRI data based on Dirichlet process mixture models'. In: *Information Processing in Medical Imaging – IPMI 2007*. Vol. 4584. LNCS. Springer, Berlin, Heidelberg, pp. 482–494. DOI: 10.1007/978-3-540-73273-0_40 (cit. on p. 64).

Tijmen Tieleman (Aug. 2013). *affNIST*. URL: https://www.cs.toronto.edu/~tijmen/affNIST/ (visited on 08/05/2018) (cit. on p. 107).

Antonio Torralba, Kevin P. Murphy, William T. Freeman and Mark A. Rubin (2003). 'Context-based vision system for place and object recognition'. In: *Proceedings Ninth IEEE International Conference on Computer Vision (ICCV 2003)*. Vol. 1, pp. 273–280. DOI: 10.1109/ICCV.2003.1238354 (cit. on pp. 86, 88).

Antonio Torralba, Alan S. Willsky, Erik B. Sudderth and William T. Freeman (2006). 'Describing visual scenes using transformed Dirichlet processes'. In: *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, pp. 1297–1304 (cit. on p. 64).

Brian L. Trippe and Richard E. Turner (2017). 'Conditional density estimation with Bayesian normalising flows'. In: *NIPS 2017 Workshop on Bayesian Deep Learning*. arXiv: 1802.04908 (cit. on pp. 131, 135).

UK Department of Health and Social Care (Sept. 2018). *Code of conduct for data-driven health and care technology*. URL: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology (visited on 10/10/2019) (cit. on p. 59).

UK National Institute for Health and Care Excellence (Mar. 2019). *Evidence Standards Framework for Digital Health Technologies*. URL: https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies (cit. on p. 59).

US Food and Drug Administration (Apr. 2019). *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)*. URL: https://www.fda.gov/media/122535/download (cit. on p. 59).

Ben Van Calster, Laure Wynants, Dirk Timmerman, Ewout W. Steyerberg and Gary S. Collins (2019). 'Predictive analytics in health care: how can we know it works?' *Journal of the American Medical Informatics Association*, ocz130. DOI: 10.1093/jamia/ocz130 (cit. on p. 59).

Cédric Villani (2009). *Optimal Transport: Old and New*. Vol. 338. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-540-71050-9 (cit. on p. 76).

Christian Wachinger and Polina Golland (2014). 'Atlas-based under-segmentation'. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*. Vol. 8673. LNCS. Springer, Cham, pp. 315–322. DOI: 10.1007/978-3-319-10404-1_40 (cit. on p. 64).

'Walking the tightrope of artificial intelligence guidelines in clinical practice' (2019). *The Lancet Digital Health* 1.3, e100. DOI: 10.1016/S2589-7500(19)30063-9 (cit. on p. 59).

Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart and Tony Yu (2014). 'scikit-image: image processing in Python'. *PeerJ* 2, e453. DOI: 10.7717/peerj.453 (cit. on p. 109).

Xiaogang Wang, W. Eric L. Grimson and Carl-Fredrik Westin (2011). 'Tractography segmentation using a hierarchical Dirichlet processes mixture model'. *NeuroImage* 54.1, pp. 290–302. DOI: 10.1016/j.neuroimage.2010.07.050 (cit. on p. 64).

Xiaogang Wang, Xiaoxu Ma and W. Eric L. Grimson (2009). 'Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models'. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.3, pp. 539–555. DOI: 10.1109/TPAMI.2008.87 (cit. on p. 99).

Thomas Wiegand, Ramesh Krishnamurthy, Monique Kuglitsch, Naomi Lee, Sameer Pujari, Marcel Salathé, Markus Wenzel and Shan Xu (2019). 'WHO and ITU establish benchmarking process for artificial intelligence in health'. *The Lancet* 394.10192, pp. 9–11. DOI: 10.1016/S0140-6736(19)30762-7 (cit. on p. 59).

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed et al. (2019). 'Do no harm: a roadmap for responsible machine learning for health care'. *Nature Medicine* 25, pp. 1337–1340. DOI: 10.1038/s41591-019-0548-6 (cit. on pp. 59, 60).

Christina Winkler, Daniel Worrall, Emiel Hoogeboom and Max Welling (2019). *Learning likelihoods with conditional normalizing flows*. arXiv: 1912.00042 (cit. on pp. 131, 135).

Herman Wold (1954). 'Causality and econometrics'. *Econometrica* 22.2, pp. 162–177. DOI: 10.2307/1907540 (cit. on p. 128).

Tian Xia, Agisilaos Chartsias, Chengjia Wang and Sotirios A. Tsaftaris (2019). *Learning to synthesise the ageing brain without longitudinal data*. arXiv: 1912.02620 (cit. on pp. 136, 150).

Han Xiao, Kashif Rasul and Roland Vollgraf (2017). *Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms*. arXiv: 1708.07747. Dataset URL: https://github.com/zalandoresearch/fashion-mnist (cit. on p. 107).

Chhavi Yadav and Léon Bottou (2019). *Cold case: the lost MNIST digits*. arXiv: 1905.10498. Dataset URL: https://github.com/facebookresearch/qmnist (cit. on pp. 105, 107).

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao and Jun Wang (2020). *CausalVAE: structured causal disentanglement in variational autoencoder*. arXiv: 2004.08697 (cit. on pp. 128, 136).

Mingde Yao, Zhiwei Xiong, Lizhi Wang, Dong Liu and Xuejin Chen (2019). 'Spectral-depth imaging with deep learning based reconstruction'. *Optics Express* 27.26, pp. 38312–38325. DOI: 10.1364/OE.27.038312 (cit. on p. 152).

Cheng Zhang, Kun Zhang and Yingzhen Li (2020). 'A causal view on robustness of neural networks'. In: *ICML 2019 Workshop on Understanding and Improving Generalization in Deep Learning*. arXiv: 2005.01095 (cit. on pp. 14, 154, 155).

Kun Zhang, Mingming Gong and Bernhard Schölkopf (2015). 'Multi-source domain adaptation: a causal view'. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2015)*. AAAI, pp. 3150–3157 (cit. on pp. 54, 56).

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet and Zhikun Wang (2013). 'Domain adaptation under target and conditional shift'. In: *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*. Vol. 28. PMLR. PMLR, pp. 819–827 (cit. on p. 54).

Le Zhang*, Ryutaro Tanno*, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Ciccarelli, Frederik Barkhof and Daniel C. Alexander (2020). 'Disentangling human error from the ground truth in segmentation of medical images'. In: *Advances in Neural Information Processing Systems (NeurIPS 2020)*. Forthcoming, *equal contribution. arXiv: 2007.15963 (cit. on p. 153).

Lei Zhang, Longbin Chen, Mingjing Li and Hongjiang Zhang (2003). 'Automated annotation of human faces in family albums'. In: *Proceedings of the Eleventh ACM International Conference on Multimedia (MULTIMEDIA '03)*. ACM Press, pp. 355–358. DOI: 10.1145/957013.957090 (cit. on p. 86).

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez and Kai-Wei Chang (2017). 'Men also like shopping: reducing gender bias amplification using corpus-level constraints'. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. ACL, pp. 2979–2989. DOI: 10.18653/v1/D17-1323 (cit. on p. 128).

Ming Zhao, Yong Wei Teo, Siliang Liu, Tat-Seng Chua and Ramesh Jain (2006). 'Automatic person annotation of family photo album'. In: *Image and Video Retrieval (CIVR 2006)*. Springer, Berlin, Heidelberg, pp. 163–172. DOI: 10.1007/11788034_17 (cit. on p. 86).

Xiaojin Zhu and Zoubin Ghahramani (2002). *Learning from Labeled and Unlabeled Data with Label Propagation*. Tech. rep. CMU-CALD-02-107. Carnegie Mellon University (cit. on p. 97).

Appendix A

Face recognition supplement

A.1 Label likelihood

Given the label prior, H_Y, we can formulate the following likelihood model:

$$F_Y(\ell \mid y_i^*) = \begin{cases} 1-\varepsilon\,, & \ell = y_i^* \\ \varepsilon \dfrac{H_Y(\ell)}{1-H_Y(y_i^*)}\,, & \ell \neq y_i^* \end{cases}. \tag{A.1}$$

Note that Eq. (A.1) depends on the unobserved label prior H_Y. Fortunately, we are able to marginalise over H_Y to obtain the following convenient result, given $\ell \neq y_i^*$:

$$\mathbb{E}\left[\frac{H_Y(\ell)}{1-H_Y(y_i^*)} \,\middle|\, \mathbf{y}^* \right] = \frac{\widehat{H_Y}(\ell \mid \mathbf{y}^*)}{1-\widehat{H_Y}(y_i^* \mid \mathbf{y}^*)}\,, \tag{A.2}$$

where $\widehat{H_Y}$ is defined as in Eq. (5.14). This straightforward equivalence arises from the fact that posterior weights in a DP follow a Dirichlet distribution and are therefore neutral: after removing one weight, the proportions between the remaining ones are independent of its value, and they simply follow a Dirichlet distribution with that component discarded.

We can then formulate an alternative likelihood, which depends on $\mathbf{y}^*$:

$$\widehat{F_Y}(\ell \mid y_i^*; \mathbf{y}^*) = \begin{cases} 1-\varepsilon\,, & \ell = y_i^* \\ \varepsilon \dfrac{\widehat{H_Y}(\ell \mid \mathbf{y}^*)}{1-\widehat{H_Y}(y_i^* \mid \mathbf{y}^*)}\,, & \ell \neq y_i^* \end{cases}. \tag{A.3}$$

Although the marginalisation in Eq. (A.2) breaks the conditional independence of the true component labels, it gives us a simple, tractable form for the likelihoods of observed labels.

The simpler case of uniform label noise, discussed in Frénay and Verleysen (2014), could not easily be extended to our context with infinite support, as this would result in an improper likelihood F_{Y}.

A.2 Gibbs sampler conditionals

We perform inference for the latent variables of our model via Gibbs sampling Markov chain Monte Carlo, i.e. by sampling each type of variable in turn from its conditional distribution given all other variables. Joint posterior over all model variables is

$$p(\mathbf{z}, \mathbf{y}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi} \,|\, \mathbf{X}, \mathbf{y}_{\mathcal{L}}, \mathbf{c}^*)\,,$$

and we derive the specific conditional distributions below.

A.2.1 Global weights

As suggested in Teh et al. (2006), we augment our Markov chain state with the weights of the global DP G_0, such that the context DPs $(G_c)_c$ become conditionally independent and can be sampled in parallel:

$$\boldsymbol{\pi}_0 = (\pi_{01}, \ldots, \pi_{0I}, \pi'_0) \,|\, \mathbf{T} \sim \mathrm{Dir}(T_{\cdot 1}, \ldots, T_{\cdot I}, \alpha_0)\,, \tag{A.4}$$

where I is the current number of distinct identities, π'_0 is the weight of G_0's base measure ($\pi'_0 = \sum_{i=I+1}^{\infty} \pi_{0i}$) and $T_{\cdot i} = \sum_{c=1}^{C} T_{ci}$ are auxiliary variables counting the total number of 'tables' (context-wise clusters) having 'dish' (global cluster) i, in the Chinese restaurant analogy (Teh et al., 2006).

Finally, to sample the table counts $\mathbf{T}$ conditioned on the global weights $\boldsymbol{\pi}$ and identity and context assignments $\mathbf{z}$ and $\mathbf{c}^*$, we use a similar scheme to the one presented in Dai and Storkey (2015):

$$T_{ci} = \sum_{n=1}^{N_{ci}} \mathbf{1}\left[u_n \leq \frac{\alpha_c \pi_{0i}}{\alpha_c \pi_{0i} + n}\right], \tag{A.5}$$

where $(u_n)_{n=1}^{N_{ci}}$ are uniformly sampled from $[0, 1]$.

A.2.2 Identity assignments

For the unlabelled instances, we have

$$p(z_n \,|\, \mathbf{X}, \mathbf{y}_{\mathcal{L}}, \mathbf{z}_{-n}, \mathbf{c}^*, \mathbf{y}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi}_0) \propto \begin{cases} F_{\mathrm{X}}(\mathbf{x}_n \,|\, \theta_i^*)\; p(z_n = i \,|\, \mathbf{z}_{-n}, \mathbf{c}^*, \boldsymbol{\pi}_0)\,, \\ \widetilde{F_{\mathrm{X}}}(\mathbf{x}_n)\; p(z_n \text{ new} \,|\, \mathbf{z}_{-n}, \mathbf{c}^*, \boldsymbol{\pi}_0)\,, \end{cases} \tag{A.6}$$

where $\widetilde{F_X}(\mathbf{x}) = \int F_X(\mathbf{x}|\theta) H_X(\theta)\,\mathrm{d}\theta$, the prior predictive distribution of the observations. The Chinese restaurant franchise conditionals $p(z_n \,|\, \mathbf{z}_{-n}, \mathbf{c}^*, \boldsymbol{\pi}_0)$ are given by (Teh et al., 2006)

$$p(z_n = i \,|\, \mathbf{z}_{-n}, \mathbf{c}^*, \boldsymbol{\pi}_0) \propto \begin{cases} N_{c_n i}^{-n} + \alpha_{c_n} \pi_{0i}\,, & N_{c_n i}^{-n} > 0 \\ \alpha_{c_n} \pi_0'\,, & i \text{ new} \end{cases}, \tag{A.7}$$

where $N_{ci} = |\{n : c_n = c \wedge z_n = i\}|$, i.e. the number of samples in context c assigned to cluster i.

The global weights $\boldsymbol{\pi}$ are updated whenever an instance gets assigned to a new cluster, by splitting π_0' according to the stick-breaking process: sample $\beta \sim \mathrm{Beta}(1, \alpha_0)$, then set $\pi_{0,I+1} \leftarrow \beta\pi_0'$ and $\pi_0' \leftarrow (1-\beta)\pi_0'$ (Teh et al., 2006).

For $n \in \mathcal{L}$, there is an additional term accounting for the likelihood of the observed label:

$$p(z_n \,|\, \mathbf{X}, \mathbf{y}_\mathcal{L}, \mathbf{z}_{-n}, \mathbf{c}^*, \mathbf{y}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi}_0) \propto F_X(\mathbf{x}_n \,|\, \theta^*_{z_n})\, \widehat{F_Y}(y_n \,|\, y^*_{z_n}; \mathbf{y}^*)\; p(z_n \,|\, \mathbf{z}_{-n}, \mathbf{c}^*, \boldsymbol{\pi}_0)\,. \tag{A.8}$$

A.2.3 Contexts

$$p(c^*_m \,|\, \mathbf{X}, \mathbf{y}_\mathcal{L}, \mathbf{z}, \mathbf{c}^*_{-m}, \mathbf{y}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi}_0) \propto p(\mathbf{z}_{\mathcal{F}_m} \,|\, \mathbf{z}_{-\mathcal{F}_m}, \mathbf{c}^*, \boldsymbol{\pi}_0)\; p(c^*_m \,|\, \mathbf{c}^*_{-m}) \tag{A.9}$$

The context posterior predictive distribution is

$$p(c^*_m = c \,|\, \mathbf{c}^*_{-m}) \propto \frac{\gamma_0}{C} + M_c^{-m}, \tag{A.10}$$

where M_c^{-m} is the number of frames assigned to context c, excluding frame m.

The conditional distribution for the identities in frame m can be computed via sequential application of Eq. (A.7):

$$p(\mathbf{z}_{\mathcal{F}_m} \,|\, \mathbf{z}_{-\mathcal{F}_m}, \mathbf{c}^*, \boldsymbol{\pi}_0) = \prod_{r=1}^{|\mathcal{F}_m|} p(z^{(r)}_{\mathcal{F}_m} \,|\, \mathbf{z}^{(<r)}_{\mathcal{F}_m}, \mathbf{z}_{-\mathcal{F}_m}, \mathbf{c}^*, \boldsymbol{\pi}_0)\,, \tag{A.11}$$

where r indexes observations within each single frame. Note that, due to exchangeability of the HDP, the order of iteration of r is inconsequential.

A.2.4 Labels

Let $J_\ell^{-i} = |\{j : y^*_j = \ell \wedge j \neq i\}|$, the number of identities with label ℓ excluding identity i, and $\mathcal{L}^{(i)} = \{n \in \mathcal{L} : z_n = i\}$, the indices of labelled observations assigned to identity i. We can then write the Gibbs

identity label predictive as

$$\widehat{H_{\mathrm{Y}}}(y_i^* \,|\, \mathbf{y}_{-i}^*) = \frac{1}{\lambda + I - 1} \begin{cases} \lambda L(\ell) + J_\ell^{-i}, & y_i^* = \ell \in \mathcal{Y} \\ \lambda(1 - L(\mathcal{Y})), & y_i^* \notin \mathcal{Y} \end{cases}, \tag{A.12}$$

where $\mathcal{Y}$ is the set of all known labels, whether allocated to components or not. Additionally, recall that the label likelihood is

$$\widehat{F_{\mathrm{Y}}}(y \,|\, \ell; \mathbf{y}^*) = (1-\varepsilon)^{\mathbf{1}[y=\ell]} \left[\varepsilon \frac{\widehat{H_{\mathrm{Y}}}(y \,|\, \mathbf{y}^*)}{1 - \widehat{H_{\mathrm{Y}}}(\ell \,|\, \mathbf{y}^*)} \right]^{\mathbf{1}[y \neq \ell]}. \tag{A.13}$$

3890 The probability of assigning a label ℓ to identity i, given the remaining identity labels, can be computed as

$$\begin{aligned} p(y_i^* = \ell \,|\, \mathbf{y}_\mathcal{L}, \mathbf{z}_\mathcal{L}, \mathbf{y}_{-i}^*) &\propto \widehat{H_{\mathrm{Y}}}(\ell \,|\, \mathbf{y}_{-i}^*) \prod_{n \in \mathcal{L}^{(i)}} \widehat{F_{\mathrm{Y}}}(y \,|\, \ell; \mathbf{y}^*) \\ &\propto \frac{\lambda L(\ell) + J_\ell^{-i}}{\lambda + I - 1} (1-\varepsilon)^{|\mathcal{L}_\ell^{(i)}|} \prod_{k \in \mathcal{Y} \setminus \{\ell\}} \left[\frac{\varepsilon(\lambda L(k) + J_k)}{\lambda + I - (\lambda L(\ell) + J_\ell)} \right]^{|\mathcal{L}_k^{(i)}|}, \end{aligned} \tag{A.14}$$

where $\mathcal{L}_\ell^{(i)} = \{n \in \mathcal{L}^{(i)} : y_n = \ell\}$.

First, let us consider the probability of assigning a *known* label to identity i:

$$\begin{aligned} &p(y_i^* = \ell \in \mathcal{Y} \,|\, \mathbf{y}_\mathcal{L}, \mathbf{z}_\mathcal{L}, \mathbf{y}_{-i}^*) \\ &\quad\propto \frac{\lambda L(\ell) + J_\ell^{-i}}{\lambda + I - 1} (1-\varepsilon)^{|\mathcal{L}_\ell^{(i)}|} \left[\frac{\varepsilon(\lambda L(\ell) + J_\ell)}{\lambda + I - (\lambda L(\ell) + J_\ell)} \right]^{-|\mathcal{L}_\ell^{(i)}|} \prod_{k \in \mathcal{Y}} \left[\frac{\varepsilon(\lambda L(k) + J_k)}{\lambda + I - (\lambda L(\ell) + J_\ell)} \right]^{|\mathcal{L}_k^{(i)}|} \\ &\quad\approx \frac{\lambda L(\ell) + J_\ell^{-i}}{\lambda + I - 1} (1-\varepsilon)^{|\mathcal{L}_\ell^{(i)}|} \left[\frac{\varepsilon(\lambda L(\ell) + J_\ell)}{\lambda + I - J_\ell} \right]^{-|\mathcal{L}_\ell^{(i)}|} \prod_{k \in \mathcal{Y}} \left[\frac{\varepsilon(\lambda L(k) + J_k)}{\lambda + I - J_\ell} \right]^{|\mathcal{L}_k^{(i)}|} \\ &\quad= \frac{\lambda L(\ell) + J_\ell^{-i}}{\lambda + I - 1} \left[\frac{(1-\varepsilon)(\lambda + I - J_\ell)}{\varepsilon(\lambda L(\ell) + J_\ell)} \right]^{|\mathcal{L}_\ell^{(i)}|} \frac{\prod_{k \in \mathcal{Y}} [\varepsilon(\lambda L(k) + J_k)]^{|\mathcal{L}_k^{(i)}|}}{(\lambda + I - J_\ell)^{|\mathcal{L}^{(i)}|}} \\ &\quad\propto \frac{\lambda L(\ell) + J_\ell^{-i}}{(\lambda + I - J_\ell)^{|\mathcal{L}^{(i)}|}} \left[\frac{(1-\varepsilon)(\lambda + I - J_\ell)}{\varepsilon(\lambda L(\ell) + J_\ell)} \right]^{|\mathcal{L}_\ell^{(i)}|}. \end{aligned} \tag{A.15}$$

where the approximation assumes that $\lambda L(\ell) \ll \lambda + I - J_\ell$, $\forall \ell$, which is generally the case for sensible choices of λ and L.

We can analogously estimate the probability of assigning an *unknown* label to an identity as follows:

$$p(y_i^* \notin \mathcal{Y} \,|\, \mathbf{y}_\mathcal{L}, \mathbf{z}_\mathcal{L}, \mathbf{y}_{-i}^*) = \sum_{\ell \notin \mathcal{Y}} p(y_i^* = \ell \,|\, \mathbf{y}_\mathcal{L}, \mathbf{z}_\mathcal{L}, \mathbf{y}_{-i}^*)$$

$$
\begin{aligned}
&\propto \sum_{\ell \notin \mathcal{Y}} \frac{\lambda L(\ell)}{\lambda + I - 1} \prod_{k \in \mathcal{Y}} \left[\frac{\varepsilon(\lambda L(k) + J_k)}{\lambda + I - \lambda L(\ell)} \right]^{|\mathcal{L}_k^{(i)}|} \\
&\approx \sum_{\ell \notin \mathcal{Y}} \frac{\lambda L(\ell)}{\lambda + I - 1} \prod_{k \in \mathcal{Y}} \left[\frac{\varepsilon(\lambda L(k) + J_k)}{\lambda + I} \right]^{|\mathcal{L}_k^{(i)}|} \\
&= \frac{\lambda(1 - L(\mathcal{Y}))}{\lambda + I - 1} \prod_{k \in \mathcal{Y}} \left[\frac{\varepsilon(\lambda L(k) + J_k)}{\lambda + I} \right]^{|\mathcal{L}_k^{(i)}|} \\
&\propto \frac{\lambda(1 - L(\mathcal{Y}))}{(\lambda + I)^{|\mathcal{L}^{(i)}|}},
\end{aligned} \tag{A.16}
$$

noting that $J_\ell = J_\ell^{-i} = |\mathcal{L}_\ell^{(i)}| = 0$ for $\ell \notin \mathcal{Y}$ and using a similar approximation as in Eq. (A.15).

Finally, combining Eqs. (A.15) and (A.16), we can summarise

$$
p(y_i^* \mid \mathbf{X}, \mathbf{y}_\mathcal{L}, \mathbf{z}, \mathbf{c}^*, \mathbf{y}_{-i}^*, \boldsymbol{\theta}^*, \boldsymbol{\pi}_0) \mathrel{\widetilde{\propto}} \begin{cases} \dfrac{\lambda L(\ell) + J_\ell^{-i}}{(\lambda + I - J_\ell)^{|\mathcal{L}^{(i)}|}} \left[\dfrac{(1-\varepsilon)(\lambda + I - J_\ell)}{\varepsilon(\lambda L(\ell) + J_\ell)} \right]^{|\mathcal{L}_\ell^{(i)}|}, & y_i^* = \ell \in \mathcal{Y} \\ \dfrac{\lambda(1 - L(\mathcal{Y}))}{(\lambda + I)^{|\mathcal{L}^{(i)}|}}, & y_i^* \notin \mathcal{Y} \end{cases}, \tag{A.17}
$$

where $\widetilde{\propto}$ means *approximately proportional to*.

A.2.5 Face feature parameters

$$
p(\theta_i^* \mid \mathbf{X}, \mathbf{y}_\mathcal{L}, \mathbf{z}, \mathbf{c}^*, \mathbf{y}^*, \boldsymbol{\theta}_{-i}^*, \boldsymbol{\pi}_0) \propto H_\mathrm{X}(\theta_i^*) \prod_{n : z_n = i} F_\mathrm{X}(\mathbf{x}_n \mid \theta_i^*), \tag{A.18}
$$

which will be analytically tractable if F_X and H_X are a conjugate pair.

Appendix B

Morpho-MNIST supplement

B.1 Morphometrics of plain and perturbed datasets

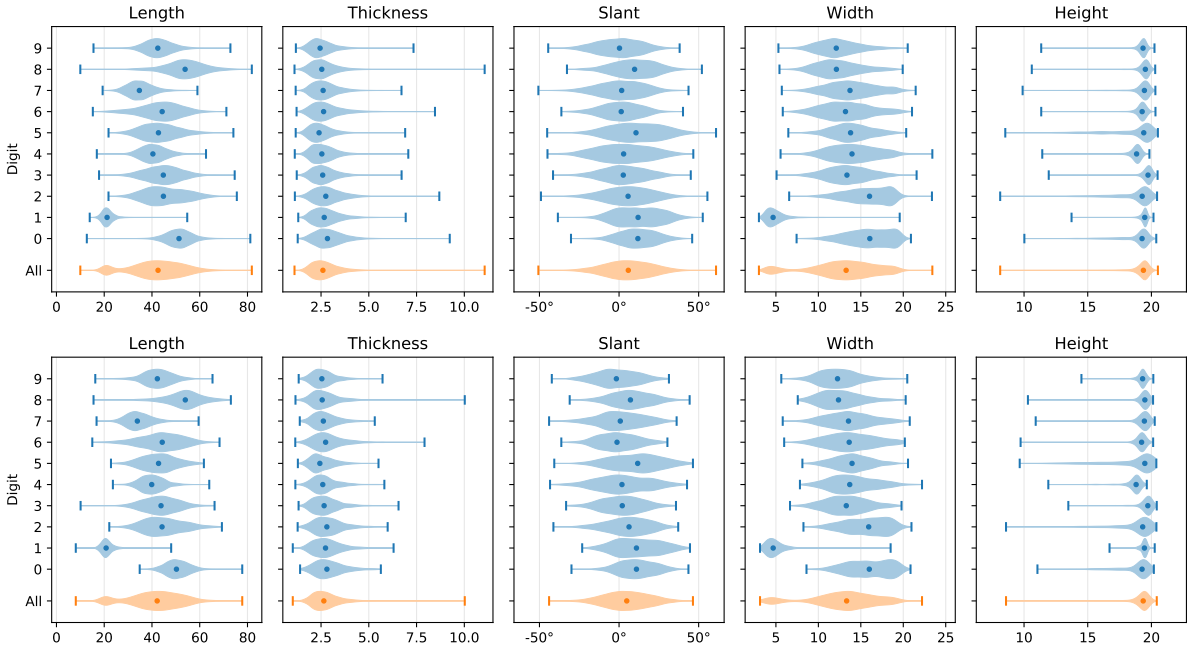


Figure B.1 Distribution of morphological attributes for plain MNIST digits. *Top:* training set; *bottom:* test set.

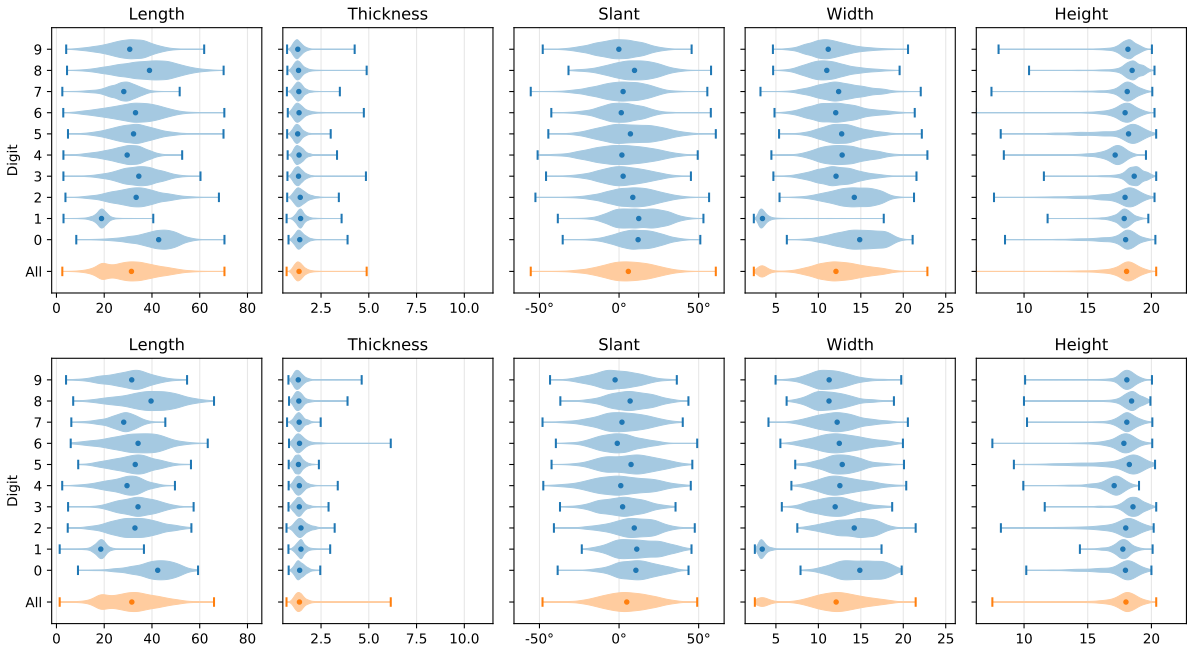


Figure B.2 Distribution of morphological attributes for thinned MNIST digits. *Top:* training set; *bottom:* test set.

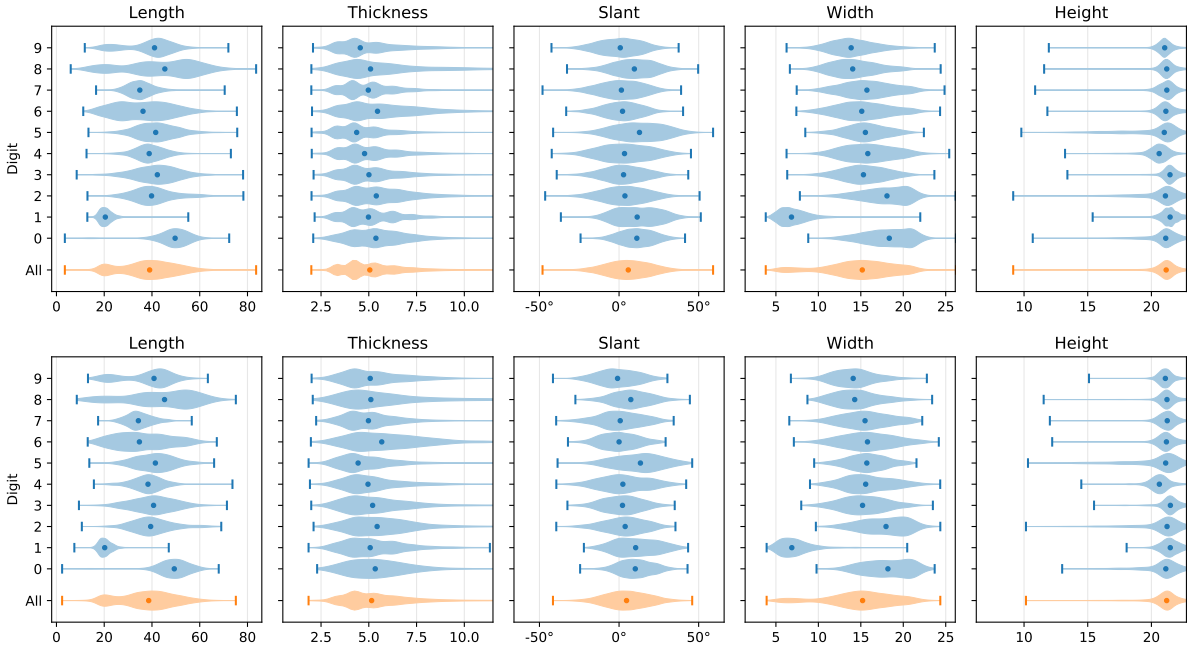


Figure B.3 Distribution of morphological attributes for thickened MNIST digits. *Top:* training set; *bottom:* test set.

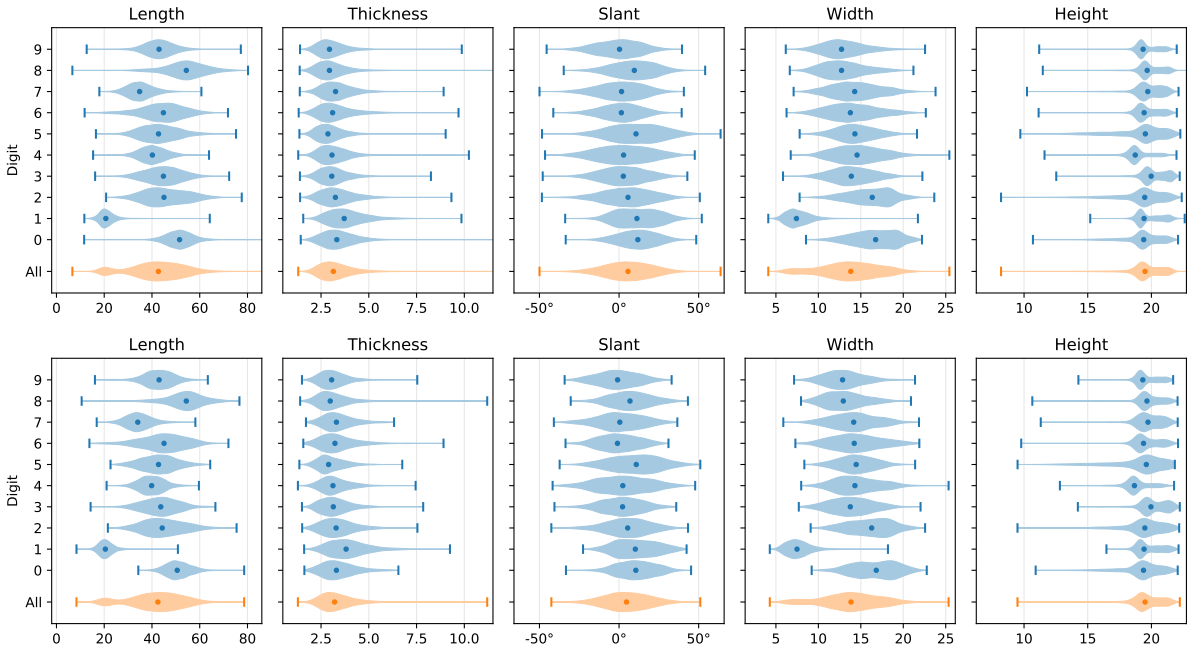


Figure B.4 Distribution of morphological attributes for swollen MNIST digits. *Top:* training set; *bottom:* test set.

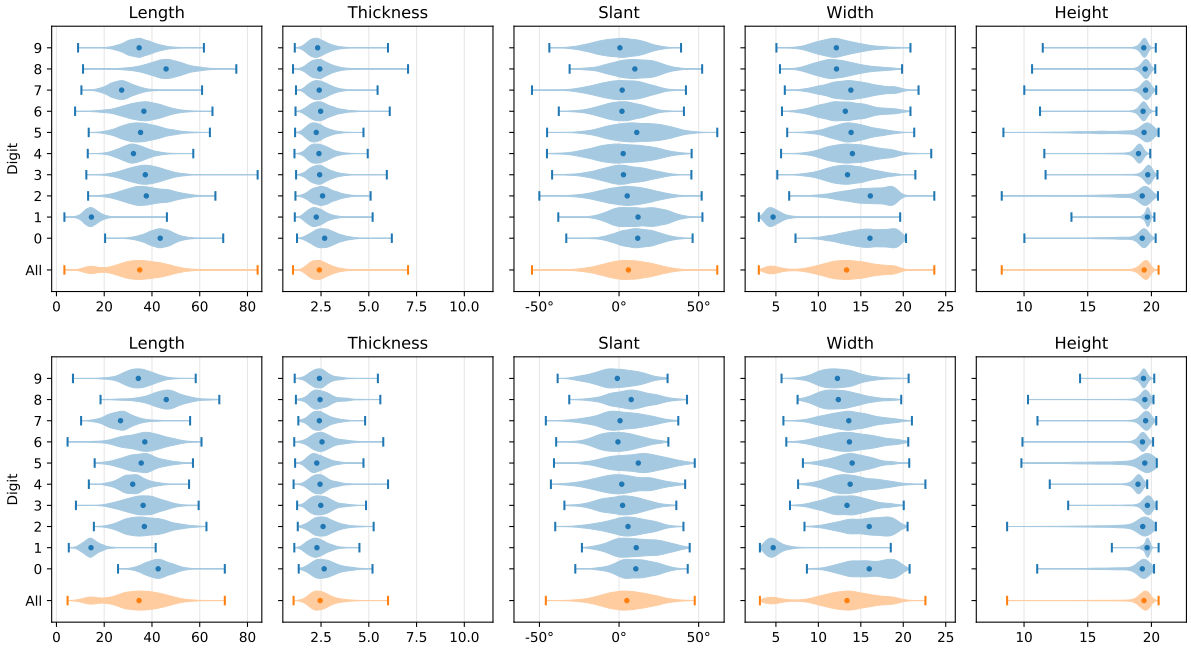


Figure B.5 Distribution of morphological attributes for fractured MNIST digits. *Top:* training set; *bottom:* test set.

B.2 Perturbation examples

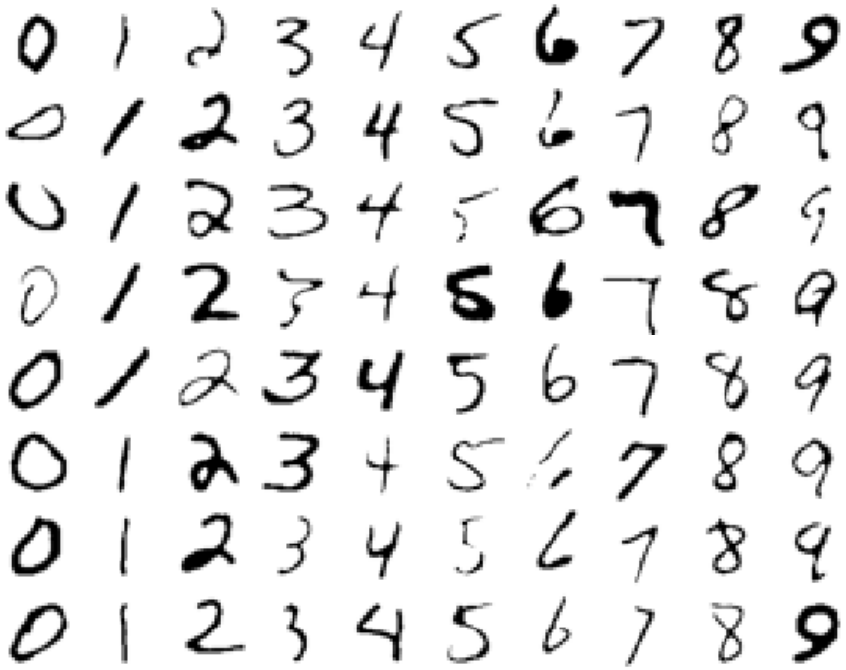


Figure B.6 Examples of globally thinned digits

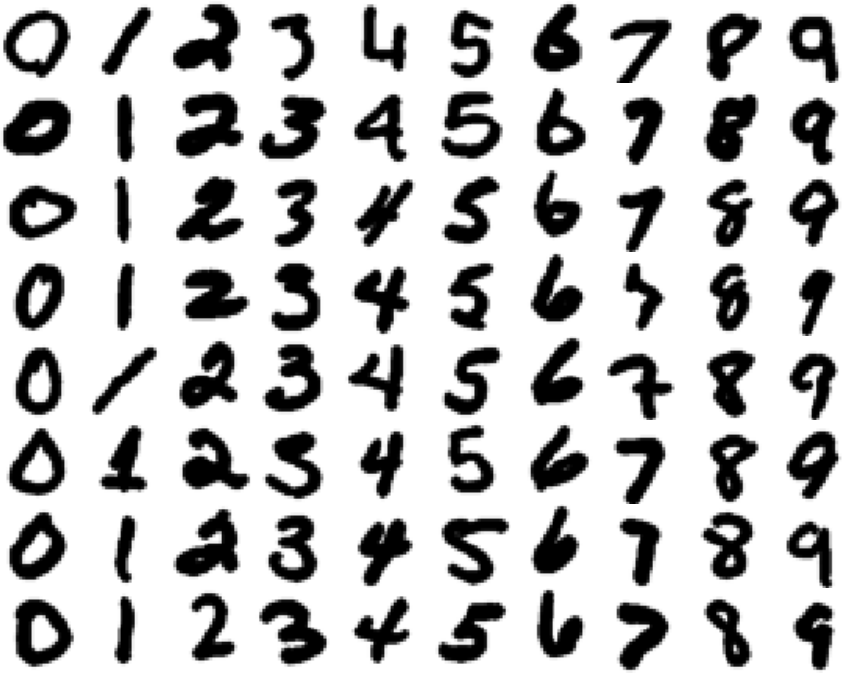


Figure B.7 Examples of globally thickened digits



Figure B.8 Examples of digits with local swelling

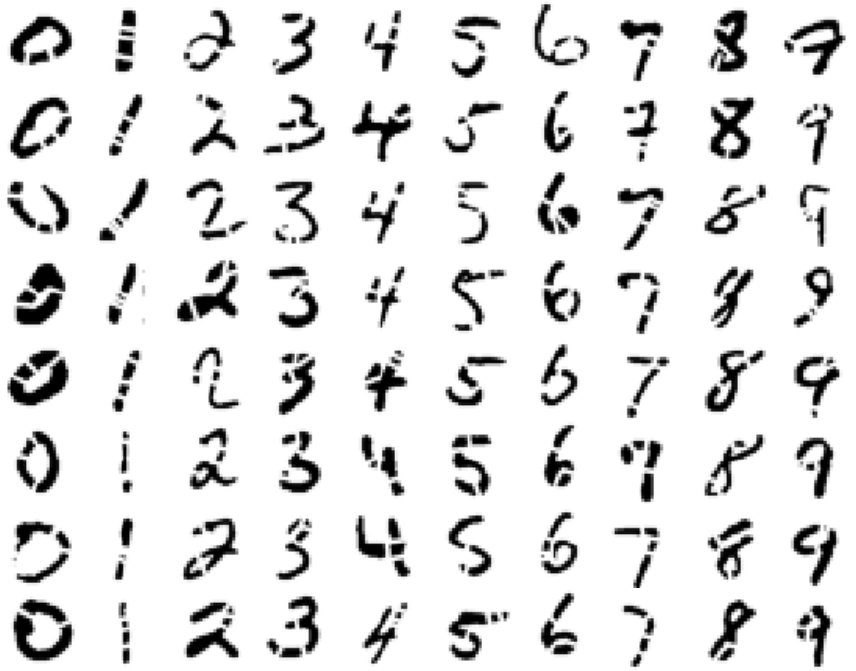


Figure B.9 Examples of digits with local fractures