

UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Probabilistic Models for Brain Image Collection, Classification, and Functional Connectivity.

Permalink

<https://escholarship.org/uc/item/0k23r2j9>

Author

Keator, David Bryant

Publication Date

2015

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA,
IRVINE

Probabilistic Models for Brain Image Collection, Classification, and Functional
Connectivity.

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

David B. Keator

Dissertation Committee:
Professor Alexander Ihler, Ph.D., Chair
Professor Padhraic Smyth, Ph.D.
Professor Steven Small, Ph.D., M.D.

2015

DEDICATION

*To my family who offered constant support and
tolerance during my graduate work.*

TABLE OF CONTENTS

	Page
LIST OF FIGURES	v
LIST OF TABLES	vi
ACKNOWLEDGMENTS	vii
CURRICULUM VITAE	ix
ABSTRACT OF THE DISSERTATION	xxiii
1 Introduction	1
1.1 Overview	3
2 Position Profile Estimation using Probabilistic Graphical Models	6
2.1 Introduction	6
2.2 Background and Related Work	8
2.2.1 HRRT Design Characteristics	9
2.2.2 Related Work	10
2.2.3 Overview of Graphical Modeling	12
2.3 Algorithm	16
2.3.1 Segmentation Model	16
2.3.2 Grid Partitioning Model	20
2.3.3 Gaussian Mixture Model	30
2.4 Experimental Results and Discussion	30
2.5 Conclusion	35
3 Feed-forward Hierarchical Model of the Ventral Visual Stream Applied to Functional Brain Image Classification	38
3.1 Introduction	38
3.2 Methods	42
3.2.1 Filtering and Feature extraction	42
3.3 Evaluation	48
3.3.1 The Alzheimer’s Disease Neuroimaging Initiative (ADNI)	48
3.3.2 AD Dataset	49
3.3.3 NFL Dataset	51
3.3.4 Ethics	52
3.3.5 Feature Sets	52
3.3.6 Classification	53

3.4	Results	55
3.5	Discussion	63
3.6	Conclusions	67
4	An Evaluation of Sparse Inverse Covariance Models for Group Functional Connectivity in Molecular Imaging	73
4.1	Introduction	73
4.2	Methods	75
4.2.1	Functional Clusters	76
4.2.2	Sparse Inverse Covariance Estimation	80
4.2.3	Gold Standard Data	85
4.3	Results	89
4.4	Discussion	99
4.5	Conclusions	101
5	Conclusion	107
5.1	Future Directions	107
	Appendices	112
A	Inverse covariance model results without clustering	113

LIST OF FIGURES

	Page
1.1 Number of ML publications over time	3
2.1 HRRT PET detector blocks	8
2.2 HRRT block amplitude differences	18
2.3 Detector center finding workflow	19
2.4 HRRT block singles	22
2.5 Detector dependency grid	23
2.6 Grid partitioning MRF	25
2.7 Model MSE with gold standard	33
2.8 Low sensitivity blocks	34
3.1 Gabor filtered PET slices	44
3.2 AD baseline classification results	61
3.3 AD 12 month classification results	62
3.4 AD 24 month classification results	63
3.5 NFL players classification results	64
4.1 Model flow chart	86
4.2 Example clustering and inverse covariance estimation	87
4.3 Gold standard SPECT cluster results	89
4.4 Gold standard male and female empirical correlation matrices	90
4.5 TP/FP for edge selection vs. gold standard $n \in \{10, 25, 50, 100\}$	94
4.6 TP/FP for edge selection vs. gold standard $n \in \{250, 500, 750, 1000\}$	95
4.7 TP/FP for edge selection vs. gold standard $n \in \{1500, 2000, 2500\}$	96
4.8 TP/FP edges by sample size and model for cross-validation experiment	98

LIST OF TABLES

	Page
3.1 Gabor filtering parameters	44
3.2 AD baseline ROC-AUC results	56
3.3 AD 12 month ROC-AUC results	57
3.4 AD 24 month ROC-AUC results	58
3.5 NFL players ROC-AUC results	59
3.6 Classification results from neuroanatomist	60
4.1 TP/FP edge area under ROC curve	93
4.2 Sum of squared errors with gold standard by model and sample size	93
4.3 Sum of squared errors with gold standard by model and sample size using cross-validation	99
A.1 TP/FP edge by model and sample size using ROI averaging	114
A.2 Sum of squared errors by model and sample size using ROI averaging	114

ACKNOWLEDGMENTS

I would like to thank Dr. Steven G. Potkin, M.D. for supporting my graduate studies and making significant sacrifices, allowing me to spend time outside of my employment to pursue my graduate studies. Without his support, I would not have been able to pursue a Ph.D.

For supporting my overall Ph.D. studies, I wish to thank Dr. Dan Cooper, M.D. and the UCI Institute for Clinical Translational Studies (ICTS). The work described in Chapters 2 and 3 were supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1 TR000153. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

For the work described in Chapter 2, I wish to thank the University of California, Irvine, Brain Imaging Center for the use of their HRRT PET scanner and Steven G. Potkin, M.D. for his support of this work.

For the work described in Chapter 3, I wish to thank the patients and healthy volunteers who participated in the studies supplying the data for this methods evaluation. This study was supported by grants to the Alzheimers Disease Neuroimaging Initiative (ADNI U01 AG024904-01), and supplement (3U01AG024904-03S5), the National Institute of Aging, the National Institute of Biomedical Imaging and Bioengineering (NIH), the Functional Imaging Biomedical Informatics Research Network (FBIRN U24-RR021992, National Center for Research Resources), the NIH through the following NCCR grant: the Biomedical Informatics Research Network (1 U24 RR025736-01). Data collection and sharing for the NFL dataset used in this study was made available by Daniel Amen, M.D. and the Amen Clinics Inc. Data collection and sharing for the AD dataset used in this project was funded by the Alzheimers Disease Neuroimaging Initiative (ADNI; Principal Investigator: Michael Weiner; NIH grant and supplement). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering (NIBIB), and through generous contributions from the following: Pfizer Inc., Wyeth Research, Bristol-Myers Squibb, Eli Lilly and Company, GlaxoSmithKline, Merck & Co. Inc., AstraZeneca AB, Novartis Pharmaceuticals Corporation, Alzheimers Association, Eisai Global Clinical Development, Elan Corporation plc, Forest Laboratories, and the Institute for the Study of Aging, with participation from the U.S. Food and Drug Administration. Industry partnerships are coordinated through the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory of Neuro-Imaging at the University of California, Los Angeles. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript (R Grant Number UL1 RR025774).

For the work described in Chapters 3 and 4, I wish to thank Dr. Daniel Amen, M.D. and the Amen Clinics Inc. for supplying the gold standard and NFL datasets. I further wish to thank members of the BrainCircuits Lab in the UCI Department of Neurology (P.I. Steven Small,

Ph.D., M.D. and Ana Solodkin, Ph.D.) for the fruitful discussions and practical advice on applying models of functional connectivity to neurology research.

CURRICULUM VITAE

David B. Keator

EDUCATION

Graduate

2008-2015 Ph.D. Computer Science - Department of Computer Science, University of California, Irvine

1997-2001 M.S. Computer Science - Department of Electrical and Computer Engineering, California State University, Long Beach

Undergraduate

1990-94 B.S. Biological Science - Department of Biological Sciences, University of California, Irvine

Language(s) English

ACADEMIC APPOINTMENTS

2013-Present Operations Director, Neuroscience Imaging Center, University of California, Irvine School of Medicine

2013-Present Chief Information Officer, BrainCircuits Lab (PIs: Steven G. Small, Ph.D., M.D., Ana Solodkin, Ph.D.), Department of Neurology, University of California, Irvine

2002-Present Specialist, Director of Scientific Computing, Brain Imaging Center, Department of Psychiatry and Human Behavior, University of California, Irvine

NON-ACADEMIC APPOINTMENTS

1997 - Present President/Founder, Brain Imaging Legal Services (www.brainimage.net)

1996 - 2002 Programmer/Analyst II, Department of Psychiatry and Human Behavior, University of California, Irvine

1994 - 1996 Staff Research Asst. I, Department of Psychiatry and Human Behavior, University of California, Irvine

1990 - 1994 EEG Technician, Department of Psychiatry and Human Behavior, University of California, Irvine

AWARDS AND HONORS

- Graduate Dean's List Award - Department of Engineering and Computer Science
- Phi Kappa Phi National Honor Society
- Ying Jow Pai Kung Fu - Black sash, Instructor
- National Eagle Scout Association

PROFESSIONAL ACTIVITY

- Review Editor - Frontiers in Brain Imaging Methods
- Reviewer - IEEE Transactions on Information Technology in Biomedicine
- Reviewer - IEEE Transactions on Medical Imaging
- Reviewer - IEEE Visualization
- Reviewer - Computer Methods and Programs in Biomedicine
- Reviewer - NeuroImage
- Reviewer - PLoS ONE

MEMBERSHIP IN PROFESSIONAL ORGANIZATIONS

- IEEE
- IEEE Computer Society

CONFERENCE PRESENTATIONS

1. Potkin SG, Kennedy J, Badri F, Jin Y, Masellis M, Gulasekaram B, Costa J, **Keator DB**, Telford J, Wu JC, Najafi A: " D1 Alleles Predict Clinical Response to sClozapine and Corresponding Brain Metabolism: A Genetic PET Scan Study," International Congress on Schizophrenia Research, 1997.
2. Potkin SG, Jin Y, Bunney B, Gulasekaram B, Costa J, **Keator DB**, Telford J, Wu JC, Najafi A, Bunney WE Jr.: " Clinical and Brain Imaging Effects of Adjunctive High Dose Glycine with Clozapine in Schizophrenia," International Congress on Schizophrenia Research, 1997.

3. Potkin SG, Arnand R, Messina J, Hartman R, **Keator D**, Wu JC, Maguire G, Fleming K, Dockstader T: FDG PET: A Sensitive and Quantitative Tool To Measure Progression of Alzheimer's Disease And Brain Metabolic Effects of Rivastigmine Tartrate -poster presented at the American Academy of Neurology, Minneapolis, MN, April 30, 1998.
4. Potkin SG, Bera R, **Keator D**, Fleming K, Alva G, Carreon D, Kranz L:" Distinguishing Predominantly Negative Symptom Schizophrenia with FDG PET." Poster presented at the International Congress on Schizophrenia Research, Santa Fe, NM, April 17-21, 1999.
5. Potkin SG, Basile VS, Badri F, **Keator D**, Wu JC, Alva G, Doo M, Bunney WE Jr., Kennedy JL: " D1 Receptor Alleles Predict PET Metabolic Correlates of Clinical Response to Clozapine." Abstract published in The International Journal of Neuropsychopharmacology, 3(Suppl 1):S6, 2000.
6. Potkin SG, Shipley J, Bera RB, Carreon D, Fallon J, Alva G, **Keator D**," Clinical and PET Effects of M100907, A Selective 5HT-2A Receptor Antagonist", abstract published in Schizophrenia Research, 49(1-2 Supplement):242.
7. Potkin, SG, Anand, R, Alva, G, Fleming, K, **Keator, D**, Carreon, D, Messina J, Wu, JC, Hartman, R, Fallon, JH: FDG PET and Clinical Effects in Placebo and Rivastigmine Treated Subjects with Alzheimer's Disease, International Congress of Geriatric Psychiatry, December 14-15, 2001, Waikola, Hawaii.
8. **Keator, D**: Medical Image Normalization and Surface Reconstruction Techniques, Youth Leadership Forum, July 10 & 24, 2002, University of California, Irvine.
9. Potkin SG, **Keator DB**, Mbgori J, Fallon JH, UC Irvine TTURC: " Brain Metabolic Effects of Nicotine Patch in High and Low Hostility Subjects," Society for Research on Nicotine & Tobacco, 2003.
10. Fallon JH, **Keator DB**, Mbgori J, Potkin SG, UC Irvine TTURC: " Gender and Ethnic Differences in Brain Metabolism Following Nicotine Patch," Society for Research on Nicotine & Tobacco, 2003.
11. Turner, J.A.; Potkin, S.G.; Brown, G.G.; Glover, G.H.; Greve, D.; **Keator, D.B.**; McCarthy, G.; Grethe, J.S.; Wible, C.G.; Lim, K.; Toga, A.W.; Andreasen, N.C.; O'Leary, D.; FIRST BIRN. Biomedical Informatics Research Network: Integrating Multi-Site Human Functional Imaging Acquisition and Analysis. Poster presented at the NIH Biomedical Information Science and Technology Initiative Consortium (BIS-TIC), 2003.

12. Potkin SG; Turner JA; Brown GG; Glover GH; Heckers S; **Keator DB**; Grethe JS. Biomedical Informatics Research Network: Functional Imaging Research in Schizophrenia Test Bed. Poster presented at Neuroscience, 2003.
13. Kemp, A.S., Aulakh, J., Jin, Y., Huerta, S.T., **Keator, D.B.**, O'Halloran, J.P., Potkin, S.G.. Fronto-Parietal EEG Coherence Correlates with Semantic Fluency and Cortical Glucose Metabolism in Patients with Alzheimer's Disease. Society of Biological Psychiatry Annual Meeting, San Francisco, May 15-17, 2003.
14. Turner, J.A.;Potkin, S.G.; Glover, G.; McCarthy, G.; Gollub, R.L.; Brown, G.; Dale, A.; Grethe, J.S.; Friedman, L.; **Keator, D.B.**; FIRST BIRN. Biomedical Informatics Research Network: Functional Imaging Research in Schizophrenia Testbed. Abstract and poster at Organization of Human Brain Mapping, 2004.
15. Ozyurt, B.I.; Wei, D.; **Keator, D.B.**; Potkin, S.G.; Brown, G.; Grethe, J.; FIRST BIRN. A General and Extensible Database System for the Storage, Retrieval and Maintenance of Human Brain Imaging and Clinical Data. Abstract and poster presentation at Organization of Human Brain Mapping, 2004.
16. Ozyurt, B.I.; Wei, D.; **Keator, D.B.**; Potkin, S.G.; Brown, G.; Grethe, J.; Web-Accessible Clinical Data Management within an Extensible Neuroimaging Database. Abstract and poster presentation at Society for Neuroscience, 2004.
17. **Keator, D.B.**; Gadde, S.; Grethe, J.S.; Taylor, D.V.; Potkin, S.G.; FIRST BIRN. A General XML Schema and Associated SPM Toolbox for Storage and Retrieval of Neuro-Imaging Results and Anatomical Labels. Abstract and poster presentation at Organization of Human Brain Mapping, 2004.
18. Fleming, K; **Keator, D.B.**; Fallon, J; Mbogori, J; Potkin, S. High Impulsivity Non-Smokers have a Robust Brain Metabolic Response to Nicotine. Abstract and poster presentation at American College of Neuropsychopharmacology (ACNP) meeting, 2004.
19. **Keator, D.B.**; Gadde, S.; Grethe, J.S.; Taylor, D.V.; Potkin, S.G.; FIRST BIRN. A General XML Schema and Associated SPM Toolbox for Storage and Retrieval of Neuro-Imaging Results and Anatomical Labels. Abstract and poster presentation at Human Brain Project Neuro-Informatics meeting, 2004.
20. **Keator, D.B.**; Gadde, S.; Grethe, J.S.; FIRST BIRN. XML Schema and Methods for a Common Image Format. Presentation at the Neuroimaging Informatics Technology Initiative (NifTI) Data Format Working Group Meeting, NIH 2004.

21. V. Sossi, H. W. A. M. de Jong, W. C. Barker, P. Bloomfield, Z. Burbar, M.-L. Camborde, R. E. Carson, C. Comtat, L. A. Eriksson, S. Houle, **D. Keator**, C. Kn, R. Kraiss, A. A. Lammertsma, A. Rahmin, M. Sibomana, M. Ters, C. J. Thompson1, R. Trbossen, J. Votaw, K. Wienhard, D. F. Wong. The second generation HRRT - a multi-centre scanner performance investigation. Abstract and poster presentation at IEEE Medical Imaging Conference, 2005.

22. B.I. Ozyurt; D. Wei; **D.B. Keator**; J.H. Bockholt; K.R. Pease; H.E. Schmidt; FIRST BIRN; Morphometry BIRN; BIRN-CC; J.S. Grethe. Scientific Data Management With an Extensible Neuroscience Database. Abstract and poster presentation at Society for Neuroscience, 2005.

23. Mukherjee J; Keator D; Collins D; Jackson H; Vivatpattanakul B; Pichika R; Christian B; Fallon J; Wu J. High Resolution HRRT-18F-Fallypride PET of the Human Brain. Abstract and poster (#652435) Radiological Society of North American Annual Meeting, 2005.

24. Potkin S; Turner J; Brown G; Stern H; Glover G; McCarthy G; Greve D; Friedman L; **Keator D**; Grethe J; Fallon J; Lim K; Gollub R; Rosen B; Toga A; Kikinis R; Lauriello J; O'Leary D; Belger A; FBIRN. Multicenter fMRI Methods and Design: Function BIRN. Abstract and poster (100.18/PP35) presentation at Society for Neuroscience, 2006.

25. **Keator D**; Ozyurt BI; Wei D; Gadde S; Potkin SG; Brown G; MBIRN; FBIRN; Grethe J. A General and Extensible Multi-Site Database and XML based Informatics System for the Storage, Retrieval, Transport, and Maintenance of Human Brain Imaging and Clinical Data. Abstract and poster (253 TH-AM) presentation at Organization of Human Brain Mapping, 2006.

26. **Keator D**; Grethe J; Ozyurt B; Gadde S; Wei D; Turner J; Potkin S; Brown G; McCarthy G; Glover G; Stern H; Lauriello J; Friedman L; Belger A; Lim K; Pieper S; Greve D; FBIRN. Function Biomedical Informatics Research Network (FBIRN) Open Source fMRI Informatics, Calibration, and Data Tools Repository. Abstract and poster (252 TH-PM) presentation at Organization of Human Brain Mapping, 2006.

27. Turner, J.; Potkin, S. G.; Brown, G. G.; Stern, H. S; Glover, G. H.; McCarthy, G.; Greve, D. N.; Friedman, L.; **Keator, D.**; Grethe, J. S. Fallon, J. H.; Lim, K. O.; Gollub, R. L.; Rosen, B. R.; Toga, A. W.; Kikinis, R.; Pieper, S.; Lauriello, J.; O'Leary, D. S.; Belger, A.; Function BIRN. Multi-center fMRI methods and design: Function BIRN. Annual Meeting of the Society for Neuroscience, Atlanta, GA. 2006.

28. Ford, J; Mathalon, D; Roach, B; Turner, J; Potkin, S; Brown, G; Wible, C; McCarthy, G; Greve, D; Mueller, B; **Keator, D**; Lim, K; O’Leary, D; Belger, A; Voyvodic, J; Glover, G; Lauriello, J; FBIRN. (2007). Hallucinations and deviant tones compete for primary auditory cortical resources: Multi-site fMRI study of schizophrenia. Annual Meeting of the Society for Neuroscience, San Diego, CA. 2007.
29. Lee K.; Hong I; Potkin S.; Burbar Z.; **Keator D**. High Resolution PET Listmode Motion Correction Using 3D Motion Data. Abstract and poster presentation at Joint Molecular Imaging Conference, 2007.
30. Lee, K.; Potkin, S.; **Keator, D.**; Hong, I. Fast 3-D Motion Correction using the Characteristics of Motion in Rigid Body. Abstract and poster presentation at IEEE Medical Imaging Conference, Dresden, Germany, 2008 .
31. Potkin, S., Turner, J., Fallon, J., Lakatos, A., **Keator, D.**, Guffanti, G., Macciardi, F., FBIRN. Gene Discovery Through Imaging Genetics: Identification of Two Novel Genes Associated with Schizophrenia. Abstract and poster presentation at International Conference on Schizophrenia Research, 2009.
32. Abe, S., Preda, A., Turner, J., **Keator, D.**, Potkin, S., FBIRN. DTI Co-registration Method Comparison Based on DTI Tractography Analysis in the Human Brain. Abstract and poster presentation at Organization of Human Brain Mapping, San Francisco, CA., 2009.
33. **Keator, D.B.**, Fowlkes, C., Fallon, J., Potkin, S., Ihler, A. Alzheimer’s Disease Classification using PET and Oriented Hierarchical Filtering. Abstract and poster presentation at Organization of Human Brain Mapping, Barcelona, Spain. 2010.
34. Bunney, W., **Keator, D.**, Potkin, S., Fallon, J., Preda, A., Mukherjee, J., Nguyen, D., Kesler-West, M., Shah, N. Ziprasidone D2 Receptor Occupancy at Doses of 120 to 240mg/day Measured with 18F-Fallypride PET Support Once-A-Day Dosing. ACNP 49th Annual Meeting, Miami Beach, FL. 2010.
35. **Keator D.B.**, Chen J., Ashish N., Torri F.,Lakatos A., Potkin S.G., Macciardi F., Wei D. HID-Genetics: A Federated BIRN-enabled Data Management System for Clinical, Imaging, and Genome-Wide Association Studies. Abstract and poster presentation at Neuroinformatics Congress 2011, Boston, MA. 2011.
36. Van Erp T.G.M, Chervenak A., Kessemann C., D’Arcy M., Sobell J., **Keator D.B.**, Dahm L., Murry J., Law M., Hasso A., Ames J., Macciardi F., Potkin S. Infrastructure for Sharing Standardized Clinical Brain Scans Across Hospitals. Paper and poster presentation at IEEE International Conference on Bioinformatics and Biomedicine, Atlanta, GA. 2011.

37. Potkin S.G., Preda A., Nguyen D., **Keator D.**, van Erp T.G.M., Kemp A., Fallon F. A brain imaging and neurocognitive study of subjects treated with aripiprazole. Abstract and poster presentation at Psych Congress, San Diego, CA 2012.
38. Helmer K., Ghosh S., Nichols N.B., **Keator D.**, Nichols T., Turner J. Connecting Brain Imaging Terms to Established Lexicons: a Precursor for Data Sharing and Querying. Abstract and poster presentation at Neuro-Informatics Congress. Munich, Germany 2012.
39. Ghosh S., Nichols N.B., Gadde S., Steffener J., **Keator D.** XCEDE-DM: A neuroimaging extension to the W3C provenance data model. Abstract and poster presentation at Neuro-Informatics Congress. Munich, Germany 2012.
40. Potkin S.G., Preda A., Nguyen D., **Keator D.B.**, Keslerwest M., Van Erp T.G.M., Kemp A., Fallon J. A Brain Imaging and Neurocognitive Study of Schizophrenic Subjects Treated with Aripiprazole. Abstract and poster presentation at U.S. Psychiatric and Mental Health Congress. San Deigo, CA 2012.
41. Nichols N., Stoner R., **Keator D.B.**, Turner J., Helmer K.G., Ashish N., Steffener J., Grabowski T.J., Ghosh S. There's an app for that: a semantic data provenance framework for reproducible brain imaging. Abstract and poster presentation at Organization of Human Brain Mapping, Seattle, WA. 2013.
42. Nichols N., Steffener J., Haselgrove C., Keator D.B., Stoner R., Poline J.B., Ghosh S. Mapping Neuroimaging Resources into the NIDASH Data Model for Federated Information Retrieval. Abstract and poster presentation at Neuroinformatics 2013, Stockholm, Sweden. 2013.
43. K.G. Helmer, S. Ghosh, **D. Keator**, C. Maumet, B.N. Nichols, T. Nichols, J.B. Poline, J. Steffener, J. Turner, W. Wong, M. Martone. The Addition of Neuroimaging Acquisition, Processing and Analysis Terms to Neurolex. Submitted abstract to Organization of Human Brain Mapping, Hamburg, Germany. 2014.
44. C. Maumet, T. Nichols, B.N. Nichols, G. Flandin, J. Turner, K.G. Helmer, J. Steffener, J.B. Poline, S. Ghosh, **D. Keator**. Extending NI-DM to share the results and provenance of a neuroimaging study: an example with SPM. Submitted abstract to Organization of Human Brain Mapping, Hamburg, Germany. 2014.
45. **D. Keator**, S. Ghosh, C. Maumet., G. Flandin, B.N. Nichols, T. Nichols, G. Burns, R. Bruehl, C. Craddock, B. Frederick, K.J. Gorgolewski, Y.O. Halchenko, M. Hanke, C. Haselgrove, K.G. Helmer, A. Klein, D. Marcus, M. Milham, F. Michel, R. Poldrack, J. Steffener, Y. Schwartz, R. Stoner, J. Turner, D. Kennedy, J.B. Poline. Developing and using the Neuroimaging and Data Sharing Data Model: the NIDASH Working Group. Abstract and Poster at Organization of Human Brain Mapping, Hamburg, Germany. 2014.

46. Wang L, Alpert KI, Calhoun V, **Keator D**, King M, Kogan A, Landis D, Tallis M, Potkin SG, Turner JA, Ambite JL. SchizConnect: Large-Scale Schizophrenia Neuroimaging Data Integration and Sharing. Annual Meeting of the American College of Neuropsychopharmacology (ACNP); December, 2011; Phoenix, Arizona, 2014.
47. Helmer K.G., Turner J. A., Maumet C., Nichols T., Nichols B.N., **Keator D.B.**, Ghosh S., Auer T., Poline J.B. Developing Terminologies for the INCF Neuroimaging Data Model (NIDM). Abstract and Poster at Organization of Human Brain Mapping, Honolulu, Hawaii. 2015.
48. Poline J.B., **Keator D.B.**, Gorgolewski K.J., Auer T., Craddock C., Flandin G., Ghosh S., Halchenko Y., Hanke M., Haselgrove C., Helmer K., Jenkinson M., Klein A., Lanyon L., Marcus D., Margulies M., Maumet C., Michel F., Nichols B.N., Nichols T., Poldrack R., Reynolds R., Saad Z., Schmah T., Steffener J., Turner J., Van Erp T., Van Horn J.D., Das S., Kennedy D. How to make brain imaging research efficient and reproducible: building software and standards. Abstract and Poster at Organization of Human Brain Mapping, Honolulu, Hawaii. 2015.
49. Nichols B.N., **Keator D.B.**, Ghosh S., Maumet C., Flandin G., Nichols T., Gorgolewski K.J., Halchenko Y.O., Hanke M., Haselgrove C., Helmer K.G., Marcus D., Poldrack R., Turner J., Kennedy D., Poline J.B., Pohl K. M. Application of the Neuroimaging Data Model to Represent and Exchange Primary and Derived Data. Abstract and Poster at Organization of Human Brain Mapping, Honolulu, Hawaii. 2015.
50. Maumet C., Nichols B.N., Flandin G., Helmer K.G., Auer T., Reynolds R., Saad Z., Chen G., Jenkinson M., Webster M., Steffener J., Gorgolewski K.J., Turner J., Nichols T., Ghosh S., Poline J.B., **Keator D.B.** Standardized reporting of neuroimaging results with NIDM in SPM, FSL and AFNI. Abstract and Poster at Organization of Human Brain Mapping, Honolulu, Hawaii. 2015.
51. **Keator D.B.**, Poline J.B., Gorgolewski K.J., Auer T., Craddock C., Flandin G., Ghosh S., Halchenko Y., Hanke M., Haselgrove C., Helmer K., Jenkinson M., Klein A., Lanyon L., Marcus D., Margulies M., Maumet C., Michel F., Nichols B.N., Nichols T., Poldrack R., Reynolds R., Saad Z., Schmah T., Steffener J., Turner J., Van Erp T., Van Horn J.D., Das S., Kennedy D. Standardizing Metadata in Brain Imaging. Submitted to Neuroinformatics Congress, Cairnes Australia, 2015.

PUBLICATIONS

Journal Articles, Peer Reviewed

1. Potkin S.G.; Wu J.; Fallon F.; Bera R.; Carreon D.; Telford J.; Plon L.; **Keator D.**; Anand R.; Hartman R. Functional neuroimaging to evaluate atypical antipsychotic compounds: an FDG PET study of SDZ MAR 327. FEBS Letters. 1995 Sep; 5(3):241-242.

2. Cahill L; Haier RJ; Fallon J; Alkire MT; Tang C; **Keator D** ; Wu J; McGaugh JL. Amygdala activity at encoding correlated with long-term, free recall of emotional information. *Proceedings of the National Academy of Sciences of the United States of America*, 1996; 93(15):8016-21.
3. Wu JC; Maguire G; Riley G; Lee A; **Keator D**; Tang C; Fallon J; Najafi A. Increased dopamine activity associated with stuttering. *Neuroreport*. 1997 Feb 10;8(3):767-70.
4. Wu JC; Bell K; Najafi A; Widmark C; **Keator D**; Tang C; Klein E; Bunney BG; Fallon J; Bunney WE. Decreasing striatal 6-FDOPA uptake with increasing duration of cocaine withdrawal. *Neuropsychopharmacology*, 1997; 17(6):402-9.
5. Wu, J; Buchsbaum MS; Gillin JC; Tang C; Cadwell S; Wiegand M; Najafi A; Hazen K; **Keator D**; Bunney WE Jr; et al. Prediction of antidepressant effects of sleep deprivation by metabolic rates in the ventral anterior cingulate and medial prefrontal cortex. *American Journal of Psychiatry*, 1999; 156(8): 1149-58.
6. Wu J; Iacono R; Ayman M; Salmon E; Lin S; Carlson J; **Keator D**; Lee A; Najafi A; Fallon J. Correlation of intellectual impairment in Parkinson's disease with FDG PET scan. *Neuroreport*, 2000; 11(10):2139-44.
7. Potkin S; Anand R; Fleming K; Alva G; **Keator D**; Carreon D; Messina J; Hartman R; Fallon J. Brain Metabolic and clinical effects of rivastigmine in Alzheimer's disease. *International Journal of Neuropsychopharmacology*, 2001; 4: 223-230.
8. Potkin S; Alva G; Fleming K; Anand R; **Keator D**; Carreon D; Doo M; Jin Y; Wu J; Fallon J. A PET study of Pathophysiology of Negative Symptoms in Schizophrenia. *American Journal of Psychiatry*, 2002; 159(2):1-11.
9. Potkin S; Alva G; **Keator D**; Carreon D; Fleming K; Fallon J. Brain metabolic effects of Neotrofin in patients with Alzheimer's disease. *Brain Research*, 2002; 951:87-95.
10. Potkin S; Basile V; Jin Y; Masellis M; Badri F; **Keator D** ; Wu J; Alva G; Carreon D; Bunney W; Fallon J; Kennedy J. D1 receptor alleles predict PET metabolic correlates clinical response to clozapine. *Molecular Psychiatry*, 2003; 8:109-113.
11. Fallon J; **Keator D**; Mbogori J; Turner J; Potkin S. Hostility differentiates the brain metabolic effects of nicotine. *Cognitive Brain Research*, 2004; 18(2):142-148.

12. Fallon J; **Keator D**; Mbogori J; Taylor D; Potkin S. Gender: a major determinant of brain response to nicotine. *International Journal of Neuropsychopharmacology*, 2004; 8:1-10.
13. Wu J; Gillin JC; Buchsbaum M; Chen P; **Keator D**; Khosla N; Darnall L; Fallon J; Bunney W. Frontal lobe metabolic decreases with sleep deprivation not total reversed by recovery sleep. *Neuropsychopharmacology*, 2006; 31(12):2783-92.
14. **Keator, D**; Gadde, S; Grethe, J; Taylor, D; FIRST BIRN; Potkin, S. A General XML Schema and Associated SPM Toolbox for Storage and Retrieval of Neuro-Imaging Results and Anatomical Labels. *Neuroinformatics*, 2006; 4(2):199-212.
15. Wu J; Gillin C; Buchsbaum B; Schachat C; Darnall L; **Keator D**; Fallon J; Bunney W. Sleep deprivation PET correlations of Hamilton symptom improvement ratings with changes in relative glucose metabolism in patients with depression. *Journal of Affective Disorders*, 2007.
16. Turner, J.A., Potkin, S.G., Brown, G.G., **Keator, D.B.**, McCarthy, G., Glover, G.H. (2007). Neuroimaging for the diagnosis and study of mental disorders. *IEEE Signal Processing Magazine*, 24(4), 112-11.
17. **Keator, D.**; Grethe, J.S.; Marcus, D.; Ozyurt, B.; Gadde, S.; Murphy, S.; Pieper, S.; Greve, D.; Notestine, R.; Bockholt, H.J; Papadopoulos, P.; Function BIRN; Morphometry BIRN; BIRNCoordinating Center. A National Human Neuroimaging Collaboratory Enabled By The Biomedical Informatics Research Network (BIRN). *IEEE Transactions on Information Technology in Biomedicine*. 2008 Mar;12(2):162-72.
18. Ford, J.; Roach, B.; Turner, J.; Brown, G.; Greve, D.; Wible, C.; McCarthy, G.; Lauriello, J.; Belger, A.; Mueller, G.; Calhoun, V.; Preda, A.; **Keator, D.**; O'Leary, D.; Lim, K.; Glover, G.; Potkin, S.; Mathalon, D. Tuning in to the voices: A multi-site fMRI study of auditory hallucinations. *Schizophrenia Bulletin*. 2009 Jan;35(1):58-66.
19. Potkin, S.G., Turner, J.A., Fallon, J.A., **Keator, D.B.** , Guffanti, G., Macciardi, F., FBIRN. Gene Discovery Through Imaging Genetics - Identification of Two Novel Genes Associated with Schizophrenia. *Molecular Psychiatry*. 2009 Apr;14(4):416-28.
20. Potkin, S.G., Turner, J.A., Guffanti, G., Lakatos, A., Torri, F., **Keator D.B.**, Macciardi F. Genome-wide strategies for discovering genetic influences on cognition and cognitive disorders: methodological considerations. *Cognitive Neuropsychiatry*. 2009;Jul;14(4):391-418.

21. Fallon, J., **Keator, D.B.** Commentary on "In silico modeling system: a national research resource for simulation of complex brain disorders." *Alzheimer's Dementia*. 2009; Jan: 5(1):5-6.
22. **Keator, D.B.**, Wei, D., Gadde, S., Bockholt, H., Grethe, J.S., Marcus, D., Aucoin, N., Ozyurt, B. Derived Data Storage and Exchange Workflow for Large-Scale Neuroimaging Analyses on the BIRN Grid. *Front Neuroinformatics*. 2009;3:30.
23. Lakatos A., Derbeneva O., Younes D., **Keator D.B.**, Bakken T., Lvova M., Brandon M., Guffanti G., Reglodi D., Saykin A., Weiner M., Macciardi F., Schork N., Wallace D., Potkin S., ADNI. Association between mitochondrial DNA variations and Alzheimer's Disease in the ADNI cohort. *Neurobiology of Aging*. 2010;31(8):1355-63.
24. Ozyurt I.B., **Keator D.**, Wei D., Fennema-Notestine C., Pease K., Bockholt B., Grethe J. Federated Web-accessible Clinical Data Management within an Extensible NeuroImaging Database. *Neuroinformatics*. 2010;23(1):98-106.
25. Amen DG, Newberg A, Thatcher R, Jin Y, Wu J, **Keator D**, Willeumier K. Impact of playing American professional football on long-term brain function. *J Neuropsychiatry Clin Neurosci*. 2011 Fall;23(1):98-106.
26. Gadde S., Aucoin N., Grethe J.S., **Keator D.B.**, Marcus D.S., Pieper S., FBIRN, MBIRN, BIRN-CC. XCEDE: An Extensible Schema for Biomedical Data. *Neuroinformatics*. 2011 Apr 9.
27. Borghammer P, Hansen SB, Eggers C, Chakravarty MM, Vang K, Aanerud JF, Hilker R, Heiss WD, Rodell A, Munk OL, **Keator D**, and Gjedde A. Glucose metabolism in small subcortical structures in Parkinson's disease. *Acta Neurol Scand*. 2011.
28. Helmer KG, Ambite JL, Ames J, Ananthakrishnan R, Burns G, Chervenak AL, Foster I, Liming L, **Keator D**, Macciardi F, Madduri R, Navarro JP, Potkin S, Rosen B, Ruffins S, Schuler R, Turner JA, Toga A, Williams C, Kesselman C; for the Biomedical Informatics Research Network. Enabling collaborative research using the Biomedical Informatics Research Network (BIRN). *J Am Med Inform Assoc*. 2011 Apr 22.
29. Rasmussen J., Lakatos A., Van Erp T., Kruggel F., **Keator D.B.**, Fallon J., Potkin S., Alzheimer's Disease Neuroimaging Initiative. Empirical derivation of the denominator region for computing degeneration sensitive ¹⁸fluorodeoxyglucose ratios in Alzheimer's Disease based on the ADNI study. *Biochim Biophys Acta - Molecular Basis of Disease*. 2012 Mar;1822(3):457-66.

30. Glover G.H., Mueller B.A., Turner J.A., Van Erp T.G, Liu T., Greve D., Voyvodic J., Rasmussen J., Brown G., **Keator D.B.**, Calhoun V.D., Lee H., Ford J., Mathalon D., Diaz M., O’Leary D., Gadde S., Preda A., Lim K., Wible C., Stern H., Belger A., McCarthy G., Ozyurt B., Potkin S.G., FBIRN. Function Biomedical Informatics Research Network Recommendations for Prospective Multi-Center Functional Magnetic Resonance Imaging Studies. *Journal of Magnetic Resonance Imaging*. 2012 Feb 7.
31. Poline J.B., Breeze J., Ghosh S., Gorgolewski K., Halchenko Y., Hanke M., Haselgrove C., Helmer K., **Keator D.B.**, Marcus D., Poldrack R., Schwartz Y., Ashburner A., Kennedy D. Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*. 2012; 6:9.
32. Chervenak A.L., van Erp T.G., Kesselman C., D’Arcy M., Sobell J., **Keator D.**, Dahm L., Murray J., Law M., Hasso A., Ames J., Macciardi F., Potkin S.G. A system architecture for sharing de-identified, research-ready brain scans and health information across clinical imaging centers. *Studies in Health Technology and Informatics*. 2012; 175:19-28.
33. **Keator D.B.**, Fallon J.H., Lakatos A., Fowlkes C., Potkin S.G., Ihler A. Feed-Forward Hierarchical Model of the Ventral Visual Stream Applied to Functional Brain Image Classification. *Journal of Human Brain Mapping*. 2012 Jul 30.
34. **Keator, D.** ”Modality Neutral Techniques for Brain Image Understanding.” *Machine Learning and Interpretation in Neuroimaging*. 2012: 84-92.
35. **Keator D.B.**, Helmer K., Steffener J., Turner J.A., Van Erp T.G.M., Gadde S., Ashish N., Burns G.A., Nichols B.N. Towards structured sharing of raw and derived neuroimaging data across existing resources. *Neuroimage*. 2013 Nov 15;82:647-61
36. Potkin, S.G., **Keator D.B.**, Kesler-West M.L., Nguyen D.D., VanErp T.G.M., Mukherjee J., Shah N., Preda A. D2 Receptor Occupancy Following Lurasidone Treatment in Patients with Schizophrenia or Schizoaffective Disorder. *CNS Spectrums*, 2013 Sep 30:1-6.
37. Raffi M.S., Baumann T.L., Bakay R.A.E, Ostrove J.M., Siffert J., Fleisher A.S., Herzog C.D., Barba D., Pay M., Tuszynski M.H., Salmon D., Kordower J.H., Bishop K., **Keator D.B.**, Potkin S.G., Bartus R.T. A phase 1 study of stereotactic gene delivery of AAV2-NGF for Alzheimer’s disease. *Alzheimer’s Disease & Dementia* 2014 Jan 7.

38. Van Erp TG, Greve DN, Rasmussen J, Turner J, Calhoun VD, Young S, Mueller B, Brown GG, McCarthy G, Glover GH, Lim KO, Bustillo JR, Belger A, McEwen S, Voyvodic J, Mathalon DH, **Keator D**, Preda A, Nguyen D, Ford JM, Potkin SG, Fbirn. A multi-scanner study of subcortical brain volume abnormalities in schizophrenia. *Psychiatry Res.* 2014;222(1-2):10-6.
39. **Keator D.B.**, Ihler, A. Position Profile Estimation Using Probabilistic Graphical Models. *IEEE Transactions on Nuclear Science.* Under Review.
40. Wang L., Alpert K., Calhoun V., **Keator D.B.**, King M., Kogan A., Landis D., Tallis M., Potkin S.G., Turner J.A., Amite J.L. SchizConnect: Mediating Schizophrenia Neuroimaging Databases for Large-Scale Integration. *Neuroimage Special Issue on Brain Imaging Repositories.* Under Review.
41. **Keator D.B.**, Van Erp T.G, Glover G.H., Mueller B.A., Turner J.A., Liu T., Greve D., Voyvodic J., Rasmussen J., Brown G., Calhoun V.D., Lee H., Ford J., Mathalon D., Diaz M., O’Leary D., Gadde S., Preda A., Lim K., Wible C., Stern H., Belger A., McCarthy G., Ozyurt B., Potkin S.G., FBIRN. Function Biomedical Informatics Research Network Brain Imaging Data Resources. *Neuroimage Special Issue on Brain Imaging Repositories.* Under Review.

Book Chapters

1. **Keator, D.B.** *Information Management in Distributed Biomedical Collaboratories.* *Methods Mol Biol: Biomedical Informatics.* 2009;569:1-23.

Illustrations, Textbook and Journal

1. " Kaplan, Sadock, et. al. Comprehensive Textbook of Psychiatry, 2000".
2. " The Journal of the California Alliance for the Mentally Ill". Weisburd, DE. Volume 8, No. 3, 1997.
3. " The California Psychologist: Neuropsychology Special Edition". Lees-Haley, P. Volume 30, No. 8, Aug 1997.

Video Publications

1. NOVA , " Can Science Stop Crime" , 2012.
2. CourtTV, " John Couey Trial" , 2007.
3. CBS News, " 48 Hour Mystery" , 2006.
4. A&E, " Love Chronicals" , 2000.
5. Discovery Magazine, " FEAR" , 1997.
6. ABC News, " The Pulse" , 1998.
7. Community Outreach, " Alliance for the Mentally Ill: Brain Imaging Center" , Mental Health Association of Orange County, 1998.
8. Discovery Magazine, " Inside the Mind of a Murderer" , 1999.

ABSTRACT OF THE DISSERTATION

Probabilistic Models for Brain Image Collection, Classification, and Functional Connectivity.

By

David B. Keator

Doctor of Philosophy in Computer Science

University of California, Irvine, 2015

Professor Alexander Ihler, Ph.D., Chair

The use of functional neuroimaging to evaluate brain disorders has become pervasive in the scientific community. The technique provides researchers with a means to evaluate dynamic in-vivo brain function. Over the last thirty years of using neuroimaging techniques to evaluate brain disorders, there is evidence suggesting some illnesses are characterized by differences in regional brain function whereas others by differences in regional connectivity. Disorders with gross anatomical and functional changes such as Alzheimer’s disease and traumatic brain injury are often visually discernible in brain scans and differences quantifiable using typical mass univariate analysis techniques. Conversely, disorders with subtle functional changes (e.g. depression) or subtle changes in how the brain communicates (e.g. schizophrenia) are less amiable to existing analysis techniques. Detecting these subtle differences in molecular imaging data, often plagued by noisy measurements from the imaging system, further impedes our ability to gain valuable insights into brain disorders. In this dissertation we use a variety of tools from machine learning and probabilistic modeling to develop new models for decreasing noise in data captured from our imaging systems, improve feature extraction for detecting differences in regional brain function, and evaluate group-based functional connectivity models and their performance in settings with small sample sizes. Each of these models are presented separately with experiments designed to show improvements over existing methodologies and measures of accuracy in both disease classification and recovering gold-standard functional relationships in the brain.

Chapter 1

Introduction

The application of machine learning methods to the field of neuroimaging has increased dramatically over the last two decades. The availability of larger data sets and increases in computational power, coupled with the growing number of machine learning techniques that perform well in settings with considerable noise and low sample sizes, relative to the number of model parameters, have contributed to this increase [4] [3] [2]. Among the myriad of tools available from the machine learning community, classification algorithms have been most heavily used, across a variety of disorders, and has become a common approach in evaluating differences in brain function [1]. More recently, the field has gained interest in evaluating how the brain communicates, leading to the application of tools from the structure learning and graph theory communities and the recent creation of the journal *Brain Connectivity* (www.liebertpub.com/overview/brain-connectivity/389/). Although the use of machine learning is a common topic at most neuroimaging conferences and in many publications, the relative amounts of available methods and literature appear to be imaging modality dependent. A search of Google Scholar (scholar.google.com) shows an exponential increase in the number of publications indexed with the terms “machine learning” and “neuroimaging” using data acquired from Magnetic Resonance Imaging (MRI) systems, yet the number of publications using data from molecular imaging systems such as Positron Emission Tomography (PET) is far lower and has slowed in recent years (Figure 1.1). Some

of this discrepancy can be explained by the greater proportion of investigators using MRI modalities as compared to PET, yet this doesn't completely explain the four-fold increase in recent publications. Another possible explanation is that molecular imaging technologies require invasive injection of radioactive isotopes and increased costs and risks relative to MRI, resulting in less available data and contributing to slower methodology development in PET relative to MRI. This reality is unfortunate because there are many benefits in using molecular imaging techniques to understand the biochemical implications of brain disorders. Approximately 27 PET tracers are in common use today, each specifically designed to measure a physiological function. Data generated from PET studies are commonly used to aid in both the understanding of a disorder and the development of pharmaceuticals to treat those disorders. Therefore, more effort should be devoted to the development of machine learning methods for molecular imaging modalities as they have the potential to improve both the clinical research that is being done and patient care.

In this dissertation we focus on developing models for molecular imaging modalities using techniques from machine learning. We begin by improving the data collected by our imaging systems. If the data collected by our instruments is poor, what we can learn from the data is limited. We develop a model to improve tuning of the imaging system, resulting in decreased noise and better spatial resolution of the collected data using techniques from probabilistic graphical models. Next, we focus on image-derived features for classification. In this work we extend a hierarchical feature model, based on theories of how the human vision system identifies objects, and apply it to finding salient features in molecular imaging data, showing the model to be highly competitive in discriminating a variety of disorders with no changes to the underlying feature computation. This model has general applicability to cases where the exact diagnosis is unknown, resulting in a lack of disorder-specific image features. Lastly, we evaluate models for group-based functional connectivity in molecular imaging, using techniques from sparse-inverse covariance estimation. In this study, we provide a comparison

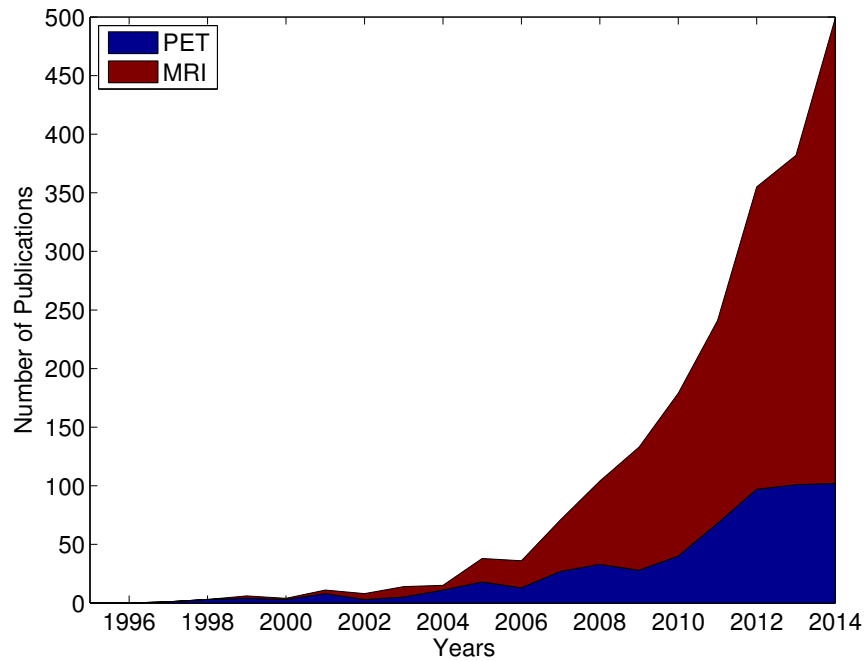


Figure 1.1: Number of publications per year indexed by Google scholar with the search terms “machine learning” and “neuroimaging” in PET and MRI modalities.

of inverse covariance models across sample sizes and evaluate their accuracies in recovering a gold standard network of functional connections in the brain. This is the first functional connectivity study in molecular imaging that shows how well we can discriminate between true functional connections and false positives by sample size and provides the community with recommendations on appropriate sample sizes and models for studying group-based functional connectivity.

1.1 Overview

This dissertation is organized into the following chapters. Each chapter is self contained and can be read independently from the others. Chapter 2 presents the work on using probabilistic graphical models to improve tuning of the imaging system. Chapter 3 presents the

work on hierarchical features and classification. Chapter 4 presents the work on functional connectivity with sparse inverse covariance estimation. Finally, Chapter 5 discusses future extensions and applications of this work.

Bibliography

- [1] Sven Haller, Karl-Olof Lovblad, Panteleimon Giannakopoulos, and Dimitri Van De Ville. Multivariate pattern recognition for diagnosis and prognosis in clinical neuroimaging: state of the art, current challenges and future trends. *Brain topography*, 27(3):329–337, 2014.
- [2] David B Keator. Modality neutral techniques for brain image understanding. In *Machine Learning and Interpretation in Neuroimaging*, pages 84–92. Springer, 2012.
- [3] Nezhir OKTAR and Yigit OKTAR. Machine learning and neuroimaging. *Journal of Neurological Sciences (Turkish)*, 32(1):001–004.
- [4] Gael Varoquaux and Bertrand Thirion. How machine learning is shaping cognitive neuroimaging. *GigaScience*, 3(1):28, 2014.

Position Profile Estimation using Probabilistic Graphical Models

2.1 Introduction

Positron emission tomography (PET) is a quantitative imaging modality used to evaluate radio-labeled tracer distributions in vivo. Superior image resolution and tracer quantification are strengths of PET over other functional imaging modalities. To ensure the imaging system is both quantitatively accurate and yielding images with maximal resolution, a manufacturer supplied suite of tools are generally used to tune the PET system. Part of the tuning process includes position profile estimation which entails locating the detection centers and boundaries of each crystal based on events collected from a gamma ray source of known activity and known position, collected over a short time interval (Figure 2.1). The accuracy of these assignments are crucial in reconstructing a quantitatively accurate estimate of the imaged object. If the assignment of detected events are incorrect, resolution will suffer and could result in misdiagnoses in clinical evaluations. Furthermore, imaging centers

pay a premium for high resolution systems and are therefore motivated to minimize setup errors that negatively impact resolution.

The High Resolution Research Tomograph (HRRT; Siemens Inc.) is one of the highest resolution PET scanners currently available and is designed for imaging the human brain [15]. There are 18 HRRT PET scanners installed worldwide. Typically the manufacturer supplied position profile estimation software makes errors in approximately 5% of the detectors. In the HRRT, this leaves about 6000 detector peaks that must be manually set by a service or site engineer. Because manually setting the peaks takes considerable time (~ 1 min/block of 64 detectors) and thus financial resources, it is of interest among the HRRT community to develop an improved position profile setup algorithm. Furthermore, the system drifts over time and the scanner tuning process must be repeated multiple times per year to ensure accurate imaging results, compounding the problem [12].

In this work, we have developed a probabilistic approach to modeling the detector center locations. The approach consists of noise segmentation followed by a grid partitioning algorithm which uses a prior over detector center configurations to constrain the search space. The algorithm is different from previous work and based on a probabilistic graphical modeling approach, which has been shown in the computer vision and machine learning communities to be efficient at modeling large systems and provides useful information about independence relationships and uncertainty. Our algorithm is fully automated, significantly outperforms the manufacturer supplied software on all our tests, and provides quantitative information about which detector settings may need manual intervention. Further, our model is general enough to be applied to many detector array configurations.

This paper describes the algorithm and the experimental results. We first present background on the HRRT detector system which motivates the design of our algorithm. Next,

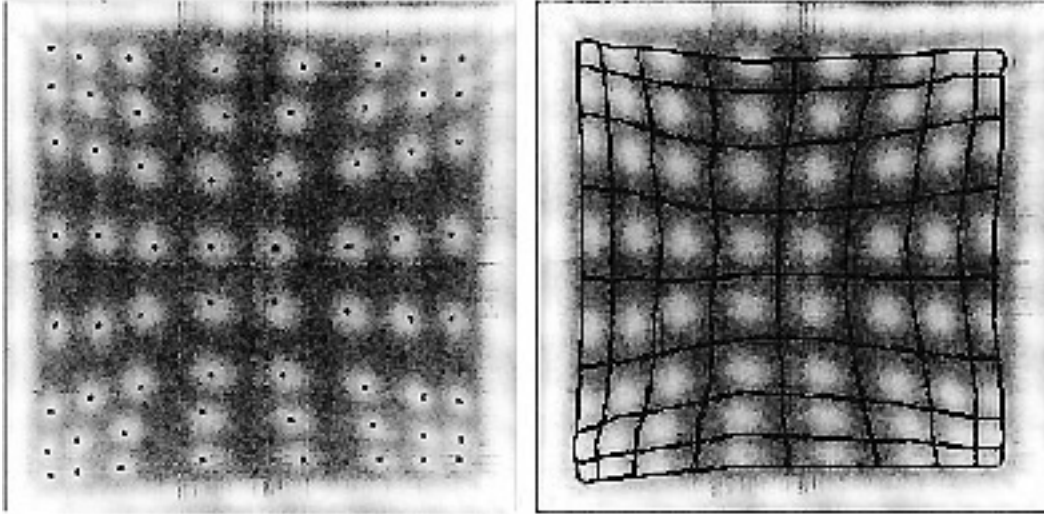


Figure 2.1: Left: HRRT PET detector block showing singles events from a 1mCi Ge-68 rod source acquired for 1 min. Crystal locations are shown with black dots. Right: Crystal borders based on crystal locations.

we discuss related work on position profile estimation in the literature. We conclude section 2.2 with a brief introduction to probabilistic graphical models and inference techniques relevant to our model. In section 2.3 we describe each component of our algorithm in detail and give the reader insight into various design considerations. In section 2.4 we present the experimental results of our algorithm compared to the manufacturer supplied software on the HRRT platform and with a typical Gaussian mixture model formulation. We conclude with a discussion of the results and plans for future work.

2.2 Background and Related Work

In this section, we briefly describe the HRRT detector system and highlight facets of the design which make the position profile estimation problem difficult. We then summarize methods found in the literature and discuss their application to the HRRT detector system. We conclude this section by introducing pairwise Markov random fields (MRFs) and inference techniques in probabilistic graphical models, useful in understanding our algorithm presented

in section 2.3.

2.2.1 HRRT Design Characteristics

The HRRT detector system is unique among PET scanner designs and is composed of 8 panels (heads) arranged in an octagon, each with a rectangular array of 119 blocks. Each detector block is 19x19x15mm and consists of two 7.5mm layers of detector material. The block is cut into an 8x8 crystal array with resulting dual layer crystals of dimensions 2.375x2.375x15mm. The HRRT PET scanner has a total of 121,856 detectors. There are no physical boundaries between detectors in each block besides cuts into the block serving as light guides. The detector block is glued to a glass plate and coupled to four photomultiplier tubes (PMTs) in a quadrant sharing design. The x and y event positions within the detector block are calculated by Anger logic [1] which uses light sharing between the four PMTs. To properly assign gamma ray detections to individual crystals, the detections for an entire block are plotted on a 256x256 pixel grid where the value of the pixel is proportional to the number of detected events at that position (Figure 2.1). The peak finding task involves locating the detection centers of each crystal based on this pixel data. The crystal locations are subsequently used to identify the crystal region boundaries in the HRRT, defined as halfway between adjacent peaks [7]. Thus in the HRRT, finding the detector centers is sufficient to obtain the crystal region boundaries. The (x, y) position of gamma ray detections are then assigned to a particular detector based on these boundaries. The manufacturer supplies a suite of tools to set up the PET system. Part of the setup process includes position profile estimation, which is done using a black box proprietary algorithm.

Because of the quadrant sharing design, areas of high amplitude noise, or crosstalk, can be seen on the borders of the block shown in Figure 2.1. Crosstalk is caused by insufficient light enclosures resulting in some light being transmitted to adjacent blocks and the quad-

rant sharing design, where events detected in neighboring blocks trigger the readout chain for all blocks linked by the shared PMT [7]. Crosstalk causes problems for automatic crystal center detection software because its presence is unpredictable before acquiring the data, the spatial distribution of the detections are similar to that of crystals within the block, and the amplitude of detections are normally much higher than the actual crystals within the block, causing problems for algorithms that search for maximal counts. The manufacturer supplied position profile setup software often mistakes the crosstalk for real detections and shifts the entire block's detector center locations into areas containing this noise. Our algorithm segments out the unwanted crosstalk and edge noise prior to modeling the detector centers, contributing, in part, to our improved results.



2.2.2 Related Work

Both parametric and non-parametric methods for position profile estimation have been proposed in the literature. In evaluating a two-dimensional array (4x8) of BGO detectors using a quadrant sharing design for high resolution PET, Dahlbom and Hoffman [4] proposed three methods for crystal identification that rely on probability distributions of x and y positions generated from a flood source phantom. Each of the three methods identify crystal boundaries based on functions of the x and y pulse height distributions. Dahlbom and Hoffman report correct detector identification rates of 76 to 87%, far too low for practical use in the HRRT system. Further, the detector block was directly coupled to the four PMTs; whereas, in the HRRT system, the four PMTs are shared by neighboring blocks, resulting in significant crosstalk, making it difficult to obtain sufficient accuracies using pulse height and position distributions. Rogers et al. [11] evaluated larger arrays (12x12 and 16x16) of BGO block detectors and also proposed an algorithm for crystal identification. Their algorithm varied

positions of vertices on a checkerboard to minimize the deviation between the expected and actual detector efficiencies. It is difficult to generalize this method because the expected detector efficiencies are needed. Furthermore it is unknown how well this method would work in the HRRT where the (x, y) positions of the detector centers in the image of the block drift spatially over time. Because we know, a-priori, the number of detectors in the block and that we expect the counts to be grouped spatially near each detector, clustering methods seem a natural fit for the problem. Xiaowen et al. [16] proposed a fuzzy c-means algorithm for 8x8 BGO crystal blocks in the MicroPET, a design similar to the HRRT. Although the authors do not give quantitative error rates, it is evident that clustering performance is poor in the corner regions where there is often no clear separation between the clusters. Furthermore, in clustering algorithms there is little control over the spatial locations of the clusters. One may find most of the clusters in a small region of the block and a few large clusters elsewhere. Stonger and Johnson [13] proposed a parametric method for crystal identification based on Gaussian mixture models and maximum likelihood estimation of mixture and model parameters. The algorithm was tested on 6x6 BGO crystal blocks, coupled to four PMTs. To initialize the mixture model, preprocessing consisted of data reduction, low pass filtering, and rules regarding deviation of counts from mean counts, peak removal heuristics, and iterative thresholding. The method presented by Stonger and Johnson relied on tuned heuristics for noise reduction to perform well and there was no discussion of how crosstalk affects the model. In comparing results from a Gaussian mixture model (GMM) with our algorithm, we find that both crosstalk and fluctuations in noise significantly affect their performance (see section 2.4).

2.2.3 Overview of Graphical Modeling

In this section we present an overview of graphical models and describe how they can be used to capture dependencies between variables. We finish the section by presenting some inference techniques that can be applied in the graph setting and are directly applicable to our algorithm. The overview in this section is intended to give the reader background in the graphical modeling and inference tools used in section 2.3. If the reader is familiar with these techniques, the section can be skipped.

2.2.3.1 Graphical Models

Probabilistic graphical models are graph-based, visual representations of joint probability distributions over variables, along with dependency relationships between those variables. Representing probability distributions as graphs are useful because they provide a simple and intuitive method of visualizing the distribution, the dependency relationships between variables can be identified by inspecting the graph, and computations to perform inference and learning can be expressed in terms of graph manipulations. A graph $G = (V, E)$ is composed of a set of vertices V , corresponding to variables in the joint probability distribution, and a set of edges E between vertices $E \subseteq \{(i, j) | i, j \in V\}$, corresponding to dependency relationships. If the edges are undirected, the graph is called a Markov network or random field (MRF) and the edges encode direct probabilistic influences between variables without enforcing a specific direction to the influence.

In undirected graphical models, the parameterization for both discrete and continuous valued random variables is through potential functions, or factors, which are functions that map an assignment of a variable (or multiple connected variables) to a positive real value $\psi(v_i, v_j) \mapsto \mathbb{R}^+$. The value returned by the potential function associated with a particular

assignment to the variable(s) generally indicates the variable's affinity for that assignment.

The most common MRF topology for computer vision and image processing is the pairwise MRF. In the pairwise MRF, each node in the network is connected to its four adjacent neighbors. Pairwise MRFs are well suited for modeling distributions where the potentials are defined over single variables (unary) or pairs of variables (pairwise). In image processing, the topology of the graph is a structured grid where nodes in the network generally correspond to pixels in the image and edges to interactions between adjacent pixels (Figure 2.5). The full joint distribution factorizes according to the graph and is the normalized product of all the unary and pairwise potentials. The domain of the variables and functional form of the potentials are determined by the image processing task. In multi-class image segmentation problems the variable's domain is generally a discrete valued region assignment and the potentials represent probabilities of configurations. Once the graph and potentials have been specified, the task in image segmentation is to infer the most probable assignment to the variables such that their joint probability is maximal.

2.2.3.2 Inference in Graphical Models

Given a graphical representation of the dependencies between variables in our joint probability distribution, our task is to answer queries using the distribution as factorized by the graphical structure. Depending on the particular query and the size and structure of the graph, we can select from a variety of exact and approximate inference algorithms designed to work efficiently in this setting. A common feature among many of these inference algorithms is the concept of message passing along edges in the graph. Messages provide a means of passing probabilistic information from one node of the graph to another and is a function associated with an edge in the graph. Depending on the query and the inference algorithm,

messages are calculated in different ways. One common query, directly applicable to our algorithm, is to find the most probable assignment to each variable (max-marginals) in the model (see section 2.3.1). To find the max-marginals, we can use the max-product message passing algorithm.

The max-product algorithm iteratively passes messages to adjacent nodes in the graph. For example, assume we have a pairwise MRF with unary potentials $\psi(v_i)$ over each node v_i and pairwise potentials $\psi(v_i, v_j)$ defined across each edge and want to calculate the message from node v_i to node v_j using the max-product update rules. The message $m_{v_i \rightarrow v_j}(v_j)$ from v_i to v_j is given by:

$$m_{v_i \rightarrow v_j}(v_j) = \max_{v_i} \left[\psi(v_i, v_j) \psi(v_i) \prod_{x \in N(v_i) \setminus v_j} m_{x \rightarrow v_i}(v_i) \right] \quad (2.1)$$

where the notation $N(v_i) \setminus v_j$ indicates the neighborhood around graph node v_i not including graph node v_j , the node in which the message is being sent. Note, the maximum operator in equation 2.1 is applied to the domain of the sending variable (v_i) making the message a function of the receiving variable (v_j). The message value is thus computed as the product of all incoming messages from adjacent nodes, not including the one receiving the message, multiplied by the unary potential of the node sending the message and the pairwise potential across the edge the message is being sent. The maximum operator is then applied to all variables in the domain of the message except the one receiving the message. In graphs that contain cycles, message values will continually change and an exact solution cannot be calculated. Instead, an approximation to the posterior distribution is calculated by iteratively passing messages between variables until the message values change by only a small amount. The final beliefs at each variable are then calculating by the product of all incoming messages from connected nodes in the graph and are the exact max-marginal probabilities in graphs without cycles. In graphs with cycles, the algorithm is not guaranteed to converge

and the beliefs are the approximate max-marginals. For more detailed information about max-product belief propagation see Wainwright and Jordan [14] and Koller and Friedman [8].

Another class of algorithms, directly applicable to our algorithm, that have been applied to graphs with cycles and in settings where exact inference algorithms are intractable due to network size and/or density of connections are the variational algorithms (see section 2.3.2). This class of approximate algorithms transforms the problem of finding the full joint distribution into an optimization problem where the approximation to the target distribution $P(V)$ takes on a simpler form. Variational algorithms reformulate the inference task into finding the maximum of an objective function defined over a set of simpler distributions $Q(V)$. The task is to find the distribution in the set that best approximates the target distribution. The simplest choice for Q is to assume that each variable in the graph is independent and represent the distribution as a product of independent marginals. The resulting algorithm is called the mean field algorithm. Assuming we have a pairwise MRF with nodes $v_{i,j}$, where (i, j) indexes a particular node in a two-dimensional matrix, and pairwise potentials $\psi(v_{i,j}, v_{k,l})$ defined across each edge, the mean field approximation for $Q(v_{i,j})$ is given by:

$$Q(v_{i,j}) = \frac{1}{Z_{i,j}} \exp \left[\begin{array}{l} \mathbb{E}_{Q(i-1,j)} [\ln (\psi(\{i-1, j\}, \{i, j\}))] + \\ \mathbb{E}_{Q(i,j-1)} [\ln (\psi(\{i, j-1\}, \{i, j\}))] + \\ \mathbb{E}_{Q(i+1,j)} [\ln (\psi(\{i+1, j\}, \{i, j\}))] + \\ \mathbb{E}_{Q(i,j+1)} [\ln (\psi(\{i, j+1\}, \{i, j\}))] + \end{array} \right] \quad (2.2)$$

where $Z_{i,j}$ normalizes the distribution. Evaluating the energy functional in mean field is therefore computed as a sum of expectations, each over small sets of connected variables. In the case of pairwise MRFs, the expectations are calculated over single variables and pairs of adjacent variables as specified by the edges in the graph. The mean field algorithm iteratively updates independent marginals (beliefs) until the marginals change by some small amount.

The algorithm is guaranteed to converge and furthermore, the distribution $\tilde{Q}(V)$ returned by mean field is guaranteed to be a fixed point of the energy functional. For more information about the mean field algorithm, see Parisi and Shankar [10], Oppen and Saad [9], and Koller and Friedman [8].

2.3 Algorithm

Using the graphical modeling approaches outlined in section 2.2.3, we now present our detector center finding algorithm. The algorithm consists of a Markov random field (MRF) based segmentation model to filter out crosstalk and noise, followed by a grid partitioning graphical model which uses candidate points and a prior over detector locations to find a configuration of detector centers. In this section we decompose the model into parts and explain each in turn.

2.3.1 Segmentation Model

One dominant characteristic of the detector block image shown in Figure 2.1, is the high amplitude crosstalk from neighboring blocks. To identify which pixels in the image are crosstalk and which are detections of interest, we developed a multiclass pairwise MRF segmentation model. Each variable in the model $v_{i,j}$ corresponds to a pixel at grid location (i, j) in the image and edges correspond to interactions between adjacent pixels in the lattice structured graph representing the image. Each variable $v_{i,j}$ has a domain $\{1, \dots, 9\}$ where the value $v_{i,j} = \alpha$ represents a region assignment for pixel (i, j) as 4 borders ($\alpha \in \{1, 3, 5, 7\}$), 4 ridges ($\alpha \in \{2, 4, 6, 8\}$), or the interior class ($\alpha = 9$) as shown in Figure 2.2. The interior segment defines the spatial location of the detections we are ultimately interested in using to model the detector centers, ridges correspond to the crosstalk, and borders to low amplitude

noise (lower than the interior signals) along the edges of the image, unlikely to be real crystal detections. The MRF is composed of node and edge potentials. The node potentials are constructed as the product of a Gaussian distributed amplitude probability and a spatial location probability and is given by:

$$\psi(v_{i,j}) = \mathcal{N}(I(i,j) | \mu_\alpha, \sigma) P(v_{i,j} = \alpha | (i,j)) \quad (2.3)$$

$$\mu_{\alpha \in \{1,3,5,7\}} = \frac{1}{N} \sum_{i,j} I(i,j) \quad (2.4)$$

$$\mu_{\alpha \in \{2,4,6,8\}} = \frac{1}{N} \sum_{i,j} I(i,j) + 2\sigma \quad (2.5)$$

$$\mu_{\alpha \in \{9\}} = \frac{1}{N} \sum_{i,j} I(i,j) + 0.25\sigma \quad (2.6)$$

The first term in equation 2.3 represents the amplitude probability which accounts for uncertainty in the detection densities ($I(i,j)$) at image pixel location (i,j) for each class α . The class dependent means μ_α of the amplitude distribution are set to the mean detection density in the image plus a scalar multiple of the standard deviation (equations 2.4, 2.5, 2.6). Reasonable settings for these means were determined empirically from previous system tunings by hand segmenting regions of the images and comparing the class means using similar imaging times and source strength. Sensible values for these would need to be determined for different source strengths and different imaging platforms. The second term in equation 2.3 represents the location probability and encourages pixels near the edge of the image to be in ridge or border segments. The location probability extents for each segment were also determined empirically from the data acquired during previous system tunings by visually estimating reasonable boundaries for each segment.

The pairwise edge potentials in the MRF penalize invalid transitions from incompatible

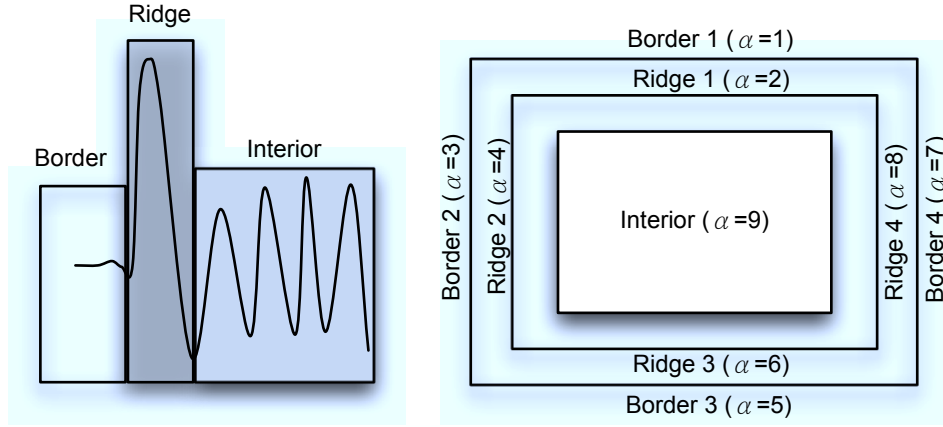


Figure 2.2: Left: Graphic depicting the amplitude differences observed between border noise, crosstalk (ridge), and detections of interest (interior) from crystals within the block. Right: Region assignment key for multiclass segmentation MRF.

segments (equation 2.7). We specify a set of constraints penalizing invalid transitions between borders, ridges, and the interior segments. In summary, invalid transitions are those from ridge to ridge or border to border that do not pass through the interior segment or are not adjacent. For example, if a variable $v_{i,j}$ is assigned the value 1 (border 1) then the edge potential penalizes the assignment of 4 (ridge2), 5 (border 3), 6 (ridge 3), and 8 (ridge 4) to adjacent variable $v_{i,j+1}$ in the graph because those assignments would skip the interior segment or are not adjacent.

$$\psi(v_{i,j}, v_{i,j+1}) = \begin{cases} 1 & \text{if } v_{i,j} = 1 \wedge v_{i,j+1} \neq \{4, 5, 6, 8\} \\ 1 & \text{if } v_{i,j} = 2 \wedge v_{i,j+1} \neq \{3, 5, 6, 7\} \\ 1 & \text{if } v_{i,j} = 3 \wedge v_{i,j+1} \neq \{2, 6, 7, 8\} \\ 1 & \text{if } v_{i,j} = 4 \wedge v_{i,j+1} \neq \{1, 5, 7, 8\} \\ \vdots & \\ \epsilon & \text{otherwise.} \end{cases} \quad (2.7)$$

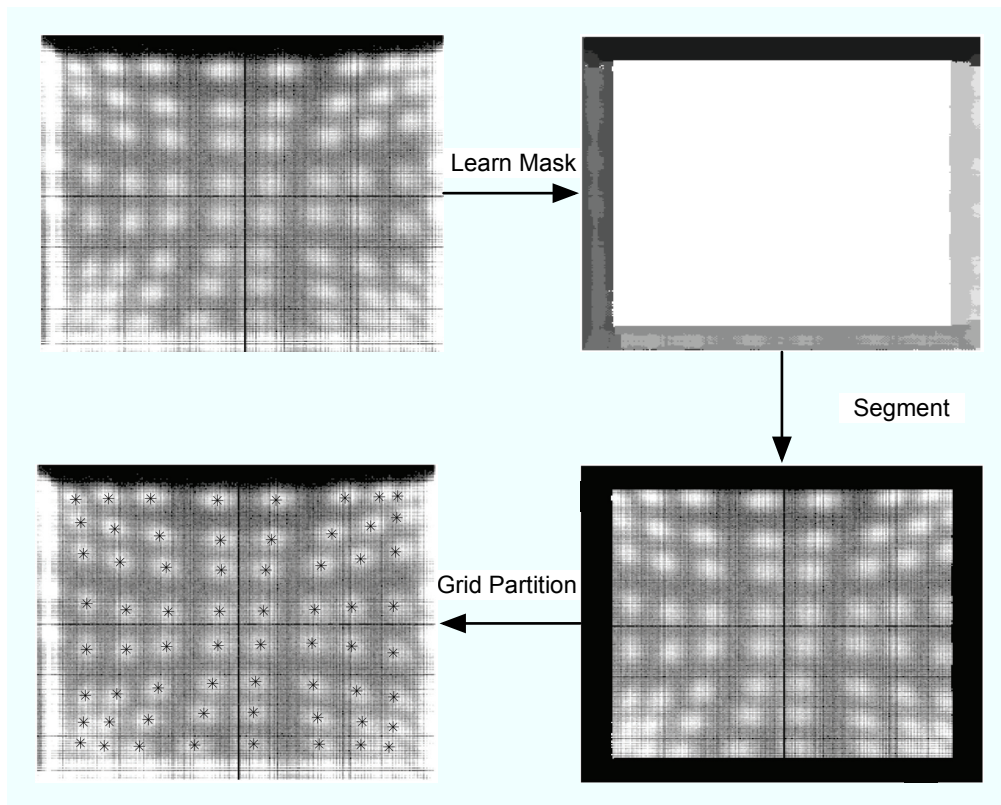


Figure 2.3: Detector center finding workflow (clockwise from upper left). Original detector image with crosstalk along three borders is used by the segmentation algorithm to learn a mask. The mask is applied to the original image masking out all segments but the interior (white central portion). The segmented detector image is used by the Grid Partitioning model to learn the location of the 64 detectors in the block.

To find the marginal probabilities for each pixel and class, we apply the max-product message passing algorithm (see section 2.2.3.2). The max-product algorithm computes the max-marginals for each variable in the graph. An image mask is generated by assigning each pixel to the class with maximum marginal probability. The mask is then used to threshold out pixels not assigned to the interior class as shown in Figure 2.3.

2.3.2 Grid Partitioning Model

Once the unwanted noise has been removed from the images using the segmentation model, the next step of our algorithm is to find the detector centers. Our model consists of a partitioning algorithm that uses candidate pixels in the image and a prior over valid configurations of candidate pixels to find the most probable configuration of detector centers. We first introduce the construction of the prior followed by the partitioning algorithm.

2.3.2.1 Configuration Prior

In reviewing configurations of detector centers in the HRRT (Figure 2.4), we observed similarities between configurations across blocks. The configurations are similar in shape yet warped with respect to each other. Further, adjacent detector centers tend to covary spatially from one block to another. To incorporate this information into our model, we define a prior distribution over detector configurations. The prior distribution is modeled as a Gaussian distribution with mean $\eta \in (\mathbb{R}^2)^m$ and covariance matrix $\Lambda^{2m \times 2m}$ where $m = 64$, corresponding to the number of detectors in each block. Each pair of elements in η is interpreted as the horizontal and vertical pixel location of the detector center for each detector. The length of the mean vector is therefore 128 elements. The covariance matrix is 128x128 elements where each 2x2 block entry $\Lambda_{ii} \in \mathbb{R}^{2 \times 2}$ for $1 \leq i \leq m$ is interpreted as the two-dimensional covariance matrix (vertical and horizontal uncertainty) for detector i 's position. The probability of a particular configuration of detector centers $d = [d_1, \dots, d_m]$ where each $d_i \in \mathbb{R}^2$ under the prior distribution is then given by:

$$P(d) = \mathcal{N}(d|\eta, \Lambda) = \frac{1}{(2\pi)^m |\Lambda|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}(d - \eta)^T \Lambda^{-1} (d - \eta) \right\} \quad (2.8)$$

The parameters η and Λ of the prior distribution are determined empirically using training data from past system configurations. The mean parameter η is calculated by averaging the detector center locations across all 936 blocks in the HRRT system. For the covariance matrix Λ , we want to capture the dependencies between adjacent detector centers and relax longer range covariances between detectors not adjacent in the block, consistent with our observations from Figure 2.4. We can visualize this network using a lattice structured grid MRF where the nodes correspond to detector centers and edges correspond to the dependencies (Figure 2.5). The lack of an edge in Figure 2.5 corresponds to a lack of dependency between the nodes. This is equivalent to 0 entries in the inverse covariance matrix Λ^{-1} of the Gaussian distribution. We therefore want to learn an inverse covariance matrix that is consistent with the empirically derived inverse covariances of detectors from the past system configurations while also having 0 entries for detectors not connected in our graphical representation.

We use an iterative proportional fitting (IPF) algorithm to learn this inverse covariance matrix [2]. The algorithm initializes Φ to the inverse empirical covariances from the training data $(\Sigma^{train})^{-1}$ for all connected nodes in the graph and 0 otherwise. Note, the entries in the inverse of the empirical covariance matrix typically contain non-zero entries between non-adjacent detectors due to the dependencies in the empirical data and/or noise. To learn entries for nodes connected in the graph that are consistent with the empirically computed covariances while keeping entries for non-adjacent detectors 0, we iteratively calculate block entries for all connected nodes as shown in equation 2.9, where block entry $\Phi_{i,j}$ is the 4x4 block of inverse covariances between detectors i and j . The notation $\Phi_{(i,j),*}$ (similarly $\Phi_{(j,i),*}$) in equation 2.9 refers to the inverse covariances of nodes i and j with all other nodes in the graph. The notation $\Phi_{*,*}$ refers to all inverse covariances except those between detectors i

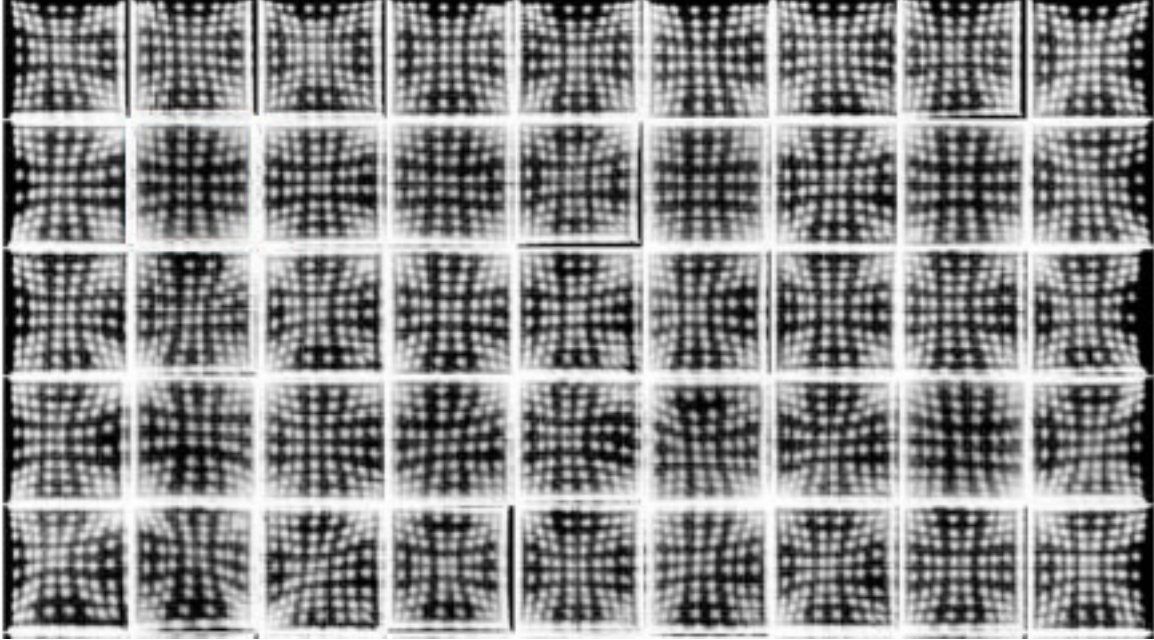


Figure 2.4: Singles events from a 1mCi Ge-68 rod source acquired for 1 min across a sub-set of blocks within one head in the HRRT system. One observes both similarities in configurations of crystals and the presence and/or absence of crosstalk on the block borders.

and j . The algorithm iterates until the inverse covariances for all connected nodes change by only some small amount ϵ , at which point the algorithm has converged. The covariance matrix Λ for our prior distribution over configurations is then set to $\Lambda = (\Phi)^{-1}$.

$$\Phi_{i,j} = (\Sigma_{i,j}^{train})^{-1} + \Phi_{(i,j),*}(\Phi_{*,*})^{-1}\Phi_{(j,i),*} \quad (2.9)$$

2.3.2.2 Grid Partitioning Model

Using the segmented detector image and the prior over configurations, our goal is to find an optimal joint distribution over detector centers such that we can use those centers to partition the detector image into a grid as shown in Figure 2.1. We approach the problem in the

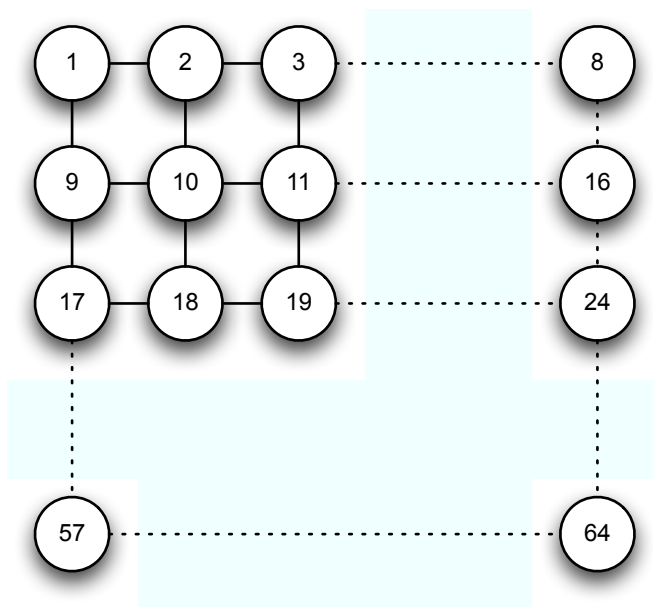


Figure 2.5: An undirected graph in which nodes correspond to detector variables and edges correspond to dependency relationships. We can read the dependencies between detector centers directly from the graph. For example, detector 1 is dependent on detectors $\{2, 9, 10\}$, detector 10 is dependent on detectors $\{2, 9, 11, 18\}$, etc.

spirit of affinity propagation [5]. In affinity propagation candidate data points are clustered by message passing, determining which of the candidate points is the best exemplar of the other points (cluster center) and which are clustered with the exemplar based on measures of responsibility and affinity respectively. Our model builds on these ideas by formulating the problem as a graphical model containing discrete candidate point variables, discrete detector variables, and continuous Gaussian random variables representing the detector centers.

The structure of the graphical model is shown in Figure 2.6. The discrete candidate point variables c_i where $i \in \{0, \dots, n\}$ indexes one of n candidate points (i.e. pixel locations) and c_0 is a null candidate point variable. Each candidate point has a spatial location in the image indicated using the notation (x_i, y_i) . The null candidate point variable has a static spatial location of $(0, 0)$. The domain of the candidate point variables is $\{0, \dots, m\}$, corresponding to the $m = 64$ detector variables plus the null ($c_i = 0$). Each of the candidate point vari-

ables are connected to each discrete detector variable r_j where $j \in \{0, \dots, m\}$, corresponding to a detector in the block plus the null detector. The domain of the detector variables is similarly $\{0, \dots, n\}$, corresponding to the n candidate points plus the null candidate point. Each detector variable r_j for $j \in \{1, \dots, m\}$ is connected to exactly one continuous Gaussian random variable d_j whose domain is the mean and variance of the Gaussian distribution over the detector center location for detector j . We introduce local unary factors $\psi_{c_i}(c_i)$ for candidate point variables and $\psi_{r_j}(r_j)$ for detector variables that penalize for not selecting a detector (i.e. $c_i = 0$) or candidate point (i.e. $r_j = 0$) respectively (equations 2.10 and 2.11).

$$\psi_{c_i}(c_i) = \begin{cases} \lambda_c & \text{if } c_i = 0 \\ 1 & \text{otherwise} \end{cases} \quad (2.10)$$

$$\psi_{r_j}(r_j) = \begin{cases} \lambda_r & \text{if } r_j = 0 \\ 1 & \text{otherwise} \end{cases} \quad (2.11)$$

To find candidate points in the image, we use the Harris corner point detector to generate a series of interest points [6]. The Harris corner point detector is invariant to rotation, scale, illumination variation and image noise. It works by measuring local changes of the signal in the image within patches that are shifted by small amounts in both the horizontal and vertical directions. The eigenvalues of a local two-dimensional neighborhood around each pixel are calculated and if the two eigenvalues are large (high variation in horizontal and vertical directions) then the pixel is on a ‘‘corner’’. Empirically we have found that detector centers roughly coincide with the Harris generated corner points; although, it finds corner points throughout the image in places other than the detector centers. We therefore obtain many more interest points than detectors in the block.

The pairwise factors between each candidate point and detector variable penalize for disagreement between variable assignments (equation 2.12). If a candidate point variable c_i 's

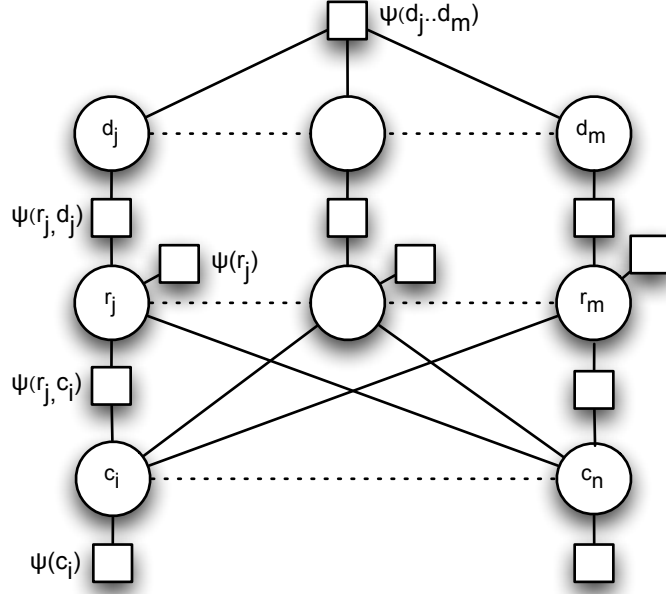


Figure 2.6: Grid partition graphical model. Variables c_i index each of the candidate points, variables r_j index detectors, and Gaussian random variables d_j represent detector center locations in the 256x256 pixel grid.

most probable assignment is j then detector variable r_j 's most probable assignment should be i . This factor encourages consistency in assignment probabilities between the variables and discourages the selection of multiple interest points and/or detectors.

$$\psi_{r_j, c_i}(r_j, c_i) = \begin{cases} \lambda_{rc} & \text{if } (r_j = i \wedge c_i \neq j) \vee (c_i = j \wedge r_j \neq i) \\ 1 & \text{otherwise} \end{cases} \quad (2.12)$$

$$\psi_{r_j, d_j}(r_j, d_j) = \begin{cases} \mathcal{N}(d_j \mid \mathbb{E}_{B(r_j)}[(x_i, y_i)], \Omega) & \text{if } r_j \neq 0 \\ \mathcal{N}(d_j \mid (0, 0), \Omega_0) & \text{if } r_j = 0 \end{cases} \quad (2.13)$$

The pairwise factor between the detector variable r_j and the corresponding Gaussian dis-

tributed detector center variable d_j (equation 2.13) indicates how likely the detector center location assigned to variable d_j is with respect to the candidate point locations (x_i, y_i) , weighted by the current belief probabilities at variable r_j for each of the candidate locations. These factors provide a way for probabilistic information to be shared between the discrete and continuous variable nodes. The variances Ω and Ω_0 in equation 2.13 are fixed and represent the uncertainty in the interest point locations and in the null detector center location respectively. The uncertainty in the interest point locations was measured empirically by generating interest points for ten images from previous system tunings and hand setting the detector peaks, then calculating the variance in Euclidean distances between the closest interest point and the hand set detector peak for each detector across all examples. The calculated standard deviation was approximately 10 pixels in the horizontal image direction and 14 pixels in the vertical image direction. The uncertainty in the null detector center (i.e. $r_j = 0$) was set to 256 pixels.

The last factor in the graphical model is the joint factor between all detector center variables d_j shown in equation 2.14. The mean η and variance Λ are the prior mean and variance over detector center configurations (see section 2.3.2.1). The joint factor evaluates to the probability of the current configuration of detector centers under the prior distribution.

$$\psi_{d_1, \dots, d_m}(d_1, \dots, d_m) = \mathcal{N}([d_1, \dots, d_m] | \eta, \Lambda) \quad (2.14)$$

Our goal is to find the detector center locations d_j for each detector. From this we can also compute the crystal region boundaries as described in section 2.2.1. By formulating our problem as a graphical model, we have implicitly defined a factorization of the full joint distribution over variables. To find the detector centers, we use the mean field algorithm (see section 2.2.3). Alternatively, MCMC sampling could be used to learn the detector centers

but would be slow to converge to the target distribution because there are many low probability regions in our model. The mean field algorithm offers a good approximate inference alternative that has lower computational costs compared to MCMC. To calculate the mean field beliefs at each variable, we formulate the updates as the expectation of all pairwise factors under the distribution formed by the product of the marginal beliefs of connected variables summed with the local unary factors.

The belief update for variable c_i is shown in equation 2.15. The first term in equation 2.15 is the expectation of the pairwise factors between each discrete detector variable r_j and the variable c_i weighted by the current beliefs of each variable r_j , where we have substituted $B(R) = \prod_{j=0}^m B(r_j)$. The second term is the local unary factor for the candidate point variable c_i . The resulting beliefs are a function of the domain of c_i and represent how well suited candidate variable c_i (i.e. pixel location) is for representing each detector center.

$$\begin{aligned}
 B(c_i) &= \exp \left[\mathbb{E}_{B(R)} [\ln \psi(r_j, c_i)] + \ln \psi(c_i) \right] \\
 &= \exp \left[\sum_{j=0}^m \sum_{k=0}^n [\ln \psi(r_j = k, c_i) B(r_j = k)] + \ln \psi(c_i) \right]
 \end{aligned} \tag{2.15}$$

The belief update for variable r_j is shown in 2.16. Similar to equation 2.15, the first term in equation 2.16 is the expectation of the pairwise factors between each candidate point variable c_i and the discrete detector variable r_j , weighted by the current beliefs of each variable c_i , where we have substituted $B(C) = \prod_{i=0}^n B(c_i)$. The second term in equation 2.16 is the probability of the current location of the detector center represented by variable d_j under a Gaussian distribution with mean centered on each candidate point and fixed variance Ω when $r_j \neq 0$ (equation 2.17). If variable $r_j = 0$ then no candidate point has been selected and the probability of the variable d_j is evaluated against the null detector center located at pixel (0,0) with variance Ω_0 . The third term in equation 2.16 is the unary factor penalizing

for not selecting a candidate point.

$$B(r_j) = \exp \left[\mathbb{E}_{B(C)} [\ln \psi(r_j, c_i)] + \mathbb{E}_{B(d_j)} [\ln \psi(r_j, d_j)] + \ln \psi(r_j) \right] \quad (2.16)$$

where

$$\mathbb{E}_{B(C)} [\ln \psi(r_j, c_i)] = \sum_{i=0}^n \sum_{k=0}^m [\ln \psi(r_j, c_i = k) B(c_i = k)]$$

and

$$\mathbb{E}_{B(d_j)} [\ln \psi(r_j, d_j)] = \begin{cases} \ln \mathcal{N}(B(d_j) \mid [x_i \ y_i], \Omega) & \text{if } ((r_j = i) \wedge (1 \leq i \leq n)) \\ \ln \mathcal{N}(B(d_j) \mid [0 \ 0], \Omega_0) & \text{if } r_j = 0 \end{cases} \quad (2.17)$$

The belief update for variable d_j is shown in equation 2.18. The first term in equation 2.18 is the expectation of the log configuration prior under the distribution formed by the product of marginal beliefs for all Gaussian detector variables $B(d_i)$, except the variable in which the belief is being calculated d_j , where we have substituted $B(d_{\forall i \setminus j}) = \prod_{\forall i \setminus j} B(d_i)$ with notation $\forall i \setminus j$ to indicate all detector variables except j . Because the expectation of a Gaussian distribution is the mean, we partition the configuration prior on the variable d_j in which the belief is being updated, separate terms involving the update variable and replace the rest with the means of the remaining variables. This forms a new Gaussian distribution show in equation 2.19 where η_j is the mean from the configuration prior for j th detector, $\Lambda_{j,*}^{-1}$ is the block inverse covariances of detector j with all other detectors from the configuration prior, μ_* are the means all variables d_i except variable d_j , and, similarly, η_* are the configuration prior means for all other detectors except j . The second term in equation 2.18 evaluates to another Gaussian distribution where the mean is the sum of all candidate points weighted by the current beliefs at variable r_j (equation 2.20). The variance is similarly the sum of the uncertainties in each candidate point, weighted by the current beliefs at variable r_j .

Together equations 2.19 and 2.20 form a new Gaussian distribution represented by variable d_j and combines the information from the current beliefs of all other detectors under the configuration prior distribution with the sum of candidate points weighted by the current beliefs of variable r_j .

$$B(d_j) \propto \exp \left[\mathbb{E}_{B(d_{\forall i \setminus j})} \left[\ln \psi_{d_{\forall i \setminus j}}(d_{\forall i \setminus j}) \right] + \mathbb{E}_{B(r_j)} [\ln \psi(r_j, d_j)] \right] \quad (2.18)$$

where

$$\mathbb{E}_{B(d_{\forall i \setminus j})} \left[\ln \psi_{d_{\forall i \setminus j}}(d_{\forall i \setminus j}) \right] = \mathcal{N}(\cdot \mid (\eta_j - \Lambda_{j,j} \Lambda_{j,*}^{-1} (\mu_* - \eta_*)), \Lambda_{j,j}) \quad (2.19)$$

$$\mathbb{E}_{B(r_j)} [\ln \psi(r_j, d_j)] \propto \mathcal{N}(\cdot \mid \theta, \Gamma) \quad (2.20)$$

and

$$\theta = \frac{\sum_{k=1}^n B(r_j = k) \Omega^{-1} [x_k \ y_k] + B(r_j = 0) \Omega_0^{-1} [0 \ 0]}{\sum_{k=1}^n B(r_j = k) \Omega^{-1} + B(r_j = 0) \Omega_0^{-1}} \quad (2.21)$$

$$\Gamma^{-1} = \sum_{k=1}^n B(r_j = k) \Omega^{-1} + B(r_j = 0) \Omega_0^{-1} \quad (2.22)$$

The mean field beliefs for each variable are calculated in an iterative fashion until the beliefs change by only a small amount. Once the algorithm has converged to tolerance, the detector centers correspond to the beliefs at each Gaussian detector center variable d_j . The detector center locations are used for comparison with our gold standard system configuration discussed in section 2.4.

2.3.3 Gaussian Mixture Model

To compare our grid partitioning model with a more typical mixture model approach similar to that proposed by Stonger and Johnson but using our configuration prior over detector centers, we introduce a Gaussian mixture model shown in equation 2.23. The two-dimensional location $([x_i \ y_i])$ of each of N detections in the detector block image (see Figure 2.1) are used to fit the mixture model. The latent random variable z indexes the mixture component, one of $m \in \{1, \dots, 64\}$ detector centers. The prior mean and variance of a valid set of detector configurations $(d = [d_1, \dots, d_m])$ are η and Λ as discussed in section 2.3.2.1. Each mixture component is a Gaussian distribution in \mathbb{R}^2 parameterized by the mean, d_m , and the covariance matrix, Σ_m , whereas the configuration prior is over all M detector centers. We use the expectation maximization (EM) algorithm to iterate between calculating the expectation of the latent random variable z using the current parameter estimates and a maximization step where the parameters are updated [3].

$$P(d, z|[X \ Y]) \propto \prod_{i=1}^N \left[\sum_{m=1}^M \mathcal{N}([x_i \ y_i] | d_m, \Sigma_m, z = m) P(z = m) \right] \mathcal{N}(d|\eta, \Lambda) \quad (2.23)$$

Because our prior is Gaussian and thus conjugate, we can maximize the mean and covariance parameters in closed form. Once the algorithm has converged to tolerance, the detector center locations are used for comparison with our gold standard system configuration discussed in section 4.3.

2.4 Experimental Results and Discussion

To evaluate the performance of our model, we used data from the HRRT system installed at the University of California, Irvine (UCI). The system was initially installed in December of

2004 and is used weekly for research and clinical PET scans. To quantitatively evaluate our model solutions, we calculate the mean squared error (MSE) rate of our model and the manufacturer supplied peak finding algorithm against a gold standard system configuration. The gold standard configuration was created by manually setting the detector center locations of 44,928 detectors in 702 detector block images. The images were acquired by scanning a $\sim 1\text{mCi}$ Ge-68 rod source for 1 minute, consistent with the manufacturer’s recommendation for turning, prior to running the manufacturer supplied setup programs which include their peak finding algorithm. For all experiments, MSE was calculated by averaging the error rates across all 702 blocks. For each block, the error rate was calculated by selecting, without replacement, the closest detector center in the model solution to the gold standard in terms of Euclidean distance, starting with detector 1 and ending with detector 64 and averaging the results.

To calculate the configuration prior, a complete system tuning from 2009 was used. Although the HRRT system is known to drift over time, we expect the parameters of the configuration prior to be relatively stable with respect to a particular HRRT machine because they are calculated over all blocks within the imaging system and no deterministic system drift has yet been identified in the HRRT literature. In our experiments, we compared our model initialized with the same previous system configuration as the manufacturer supplied software. To further evaluate our segmentation and grid partitioning model (Seg+GridPart), we compare it with the Gaussian mixture model presented in section 2.3.3. We compare with the mixture model in three ways. First, the mixture model is run using the un-segmented data and the default system configuration as the starting point (GMM). Second, the mixture model is run after applying the segmentation model and using the same default system configuration (Seg+GMM). Finally, the mixture model is run after applying the segmentation model and initializing with the results from the grid partitioning model (Seg+GridPart+GMM). For each solution we calculate the MSE against the gold standard configuration.

Using our segmentation+grid partitioning (Seg+GridPart) model we obtained a statistically significant improvement in the error rates over the manufacturer supplied software. The graph in Figure (2.7) shows the log mean squared error rate in estimating the peak locations over the 702 detector blocks as compared to the gold standard configuration. Our Seg+GridPart model yielded a 39% improvement ($t = 4.25, p < 0.008$) in MSE rate over the manufacturer supplied software. Both the Seg+GMM and Seg+GridPart+GMM models yielded an 18% improvement ($t = 2.25, p < 0.08$) over the Siemens model. When adding the GMM model, both initializing with the default system configuration or the GridPart solution results in virtually identical MSE rates due to the GMM algorithm’s necessity to maximize the likelihood of all the data under the mixture model. The GMM model alone on un-segmented data performed much worse than the others, caused by the presence of significant crosstalk in the un-segmented images and the likelihood properties of the GMM model in the EM framework mentioned above.

To evaluate our model in the presence of low count rates, we acquired images using a $0.57\mu\text{Ci}$ Ge-68 rod source for 1 minute. We performed the same experiment as described previously, computing the MSE error rate of our model with the manually set gold standard system configuration. The manufacturer supplied setup programs were not stable using the low activity source and could not be used to tune the system. Our Seg+GridPart model, even in the low count environment, yielded a 21% improvement, on average, in MSE compared to manufacturer’s model acquired with a 1mCi source; whereas, the Seg+GridPart+GMM model yielded a 6% improvement. These results suggest our model still performs well with low count rates.

In our experiments, our model significantly outperforms the manufacturer supplied peak-finding software. In evaluating errors made by our model compared to those made by the

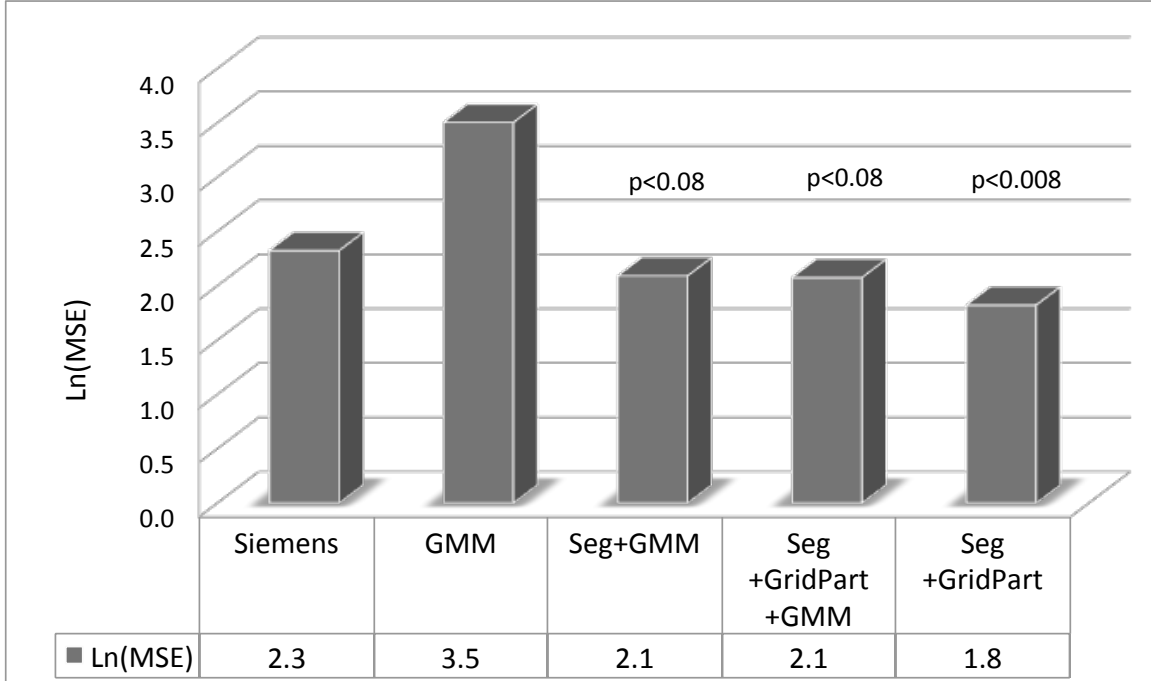


Figure 2.7: Log mean squared error rates of our Grid Partitioning with noise segmentation model (Seg+GridPart) versus the Siemens’ peak finding model shows a 39% decrease ($t = 4.25, p < 0.008$) in MSE. Our full model is compared to the Gaussian mixture model (GMM) without noise segmentation, the Gaussian mixture model with segmentations (Seg+GMM), and the Gaussian mixture model with segmentations and initialized with the results from the GridPart model (Seg+GridPart+GMM) versus the gold standard configuration across 702 detector blocks (44,928 detectors). All models were initialized using the default 2004 system configuration to be consistent with the Siemens tuning software unless otherwise indicated.

manufacturer’s, we find that our model is typically making errors in the detectors at the corners of the blocks. In the corner regions there is much more variability in block to block detector configurations. Another source of error in our model solutions stem from the Harris corner point detector. Our GridPart model relies on having reasonable candidate points available. If the Harris corner point algorithm does not generate candidate points near detector centers, the accuracy of the model results will suffer. The interaction between the quality of the segmentations and the corner points generated further affects our algorithm’s performance; although, we have found the segmentation model to be reliable even in low count environments. The manufacturer’s software generally makes errors in both the corners of

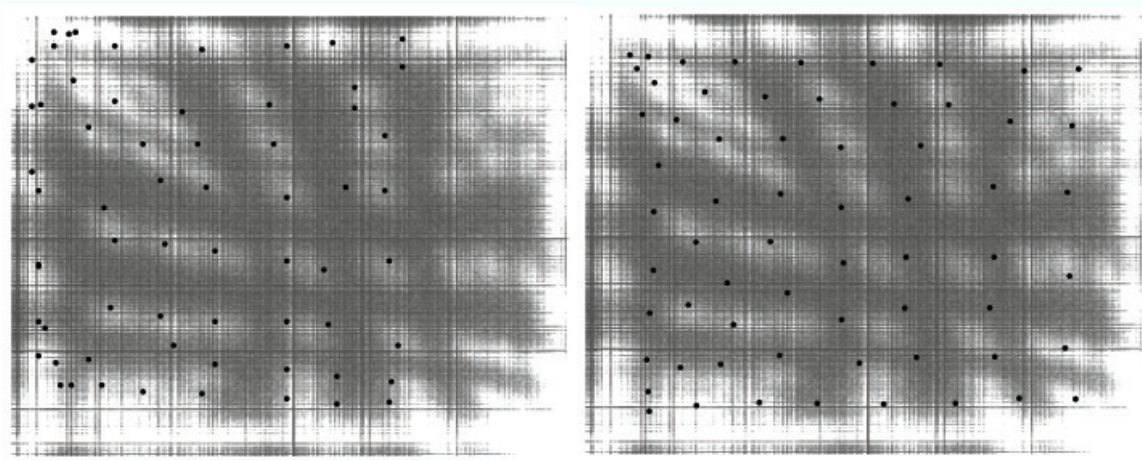


Figure 2.8: Block with unclear detector boundaries and low sensitivity. Our Seg+GridPart model (right) gets closer to the correct configuration than the manufacturer’s peak finding software (left).

the blocks and shifting errors where the entire block configurations are shifted into areas of crosstalk. These shifting errors, common in the manufacturer’s solution, while also present in the un-segmented GMM model results, take more time to manually fix than adjusting a few misplaced detector centers because each detector center must be either moved from its current location or reset from scratch. To completely re-set all the detector centers in one HRRT detector block using the manufacturer’s supplied software it takes, on average, one minute compared to a few seconds to adjust the location of a detector center. Typically after a full system setup, an engineer must manually review the configurations for all blocks to evaluate accuracy. This process takes considerable time. Using our model, we can evaluate the likelihood of each block configuration under the prior distribution over configurations and rank order blocks that have low likelihood, giving engineers guidance on what blocks in the system may need manual intervention. Another benefit from our model is in blocks with low counts across the block and/or in blocks where detector boundaries are unclear (Figure 2.8). In these situations, our model draws more heavily on information from the prior distribution over configurations to yield more sensible results which are closer to the true configuration. Comparing the image on the right of Figure 2.8 with the one on the left,

the black dots are closer to their respective detector centers. In the image on the left, there are many detector centers that end up in the crosstalk regions both to the left and at the top of the panel. The overall relative relationship between the detector centers in the left image are not consistent with the shape of the block whereas the ones on the right are more consistent. Lastly, the MSE between the gold standard configuration for this block is lower in our model results than the manufacturers result on the left.

2.5 Conclusion

In this work we have developed a probabilistic approach to position profile estimation in PET detector systems and applied it to the detector arrays in the HRRT. Our model consists of a system specific prior over detector configurations, a noise segmentation algorithm, and a grid partitioning algorithm. The model is general enough to be used on many detector configurations. The model was used to successfully estimate a position profile on the HRRT PET system. Our model outperforms the manufacturer’s supplied position profile software with a 39% drop in mean squared error, on average, in our tests. Further, it performs better on panels with low sensitivity and/or non-standard detector configurations and requires no manual intervention to run once the prior distributions have been computed for the specific PET system. In settings with low count rates, our model still outperforms the manufacturer’s software using a much stronger source. Future work will consist of both deploying the solution on other HRRT systems and developing quality control extensions, allowing engineers to track changes in the position profile estimates over time. In addition, we are interested in applying the model to other PET scanner block configurations and detector materials to further evaluate its performance.

Bibliography

- [1] H.O. Anger. Scintillation camera. *Review of Scientific Instruments*, 29(1):27–33, 1958. ISSN 0034-6748.
- [2] Michael Bacharach. Estimating nonnegative matrices from marginal data. *International Economic Review*, 6(3):294–310, 1965.
- [3] C.M. Bishop et al. *Pattern recognition and machine learning*. Springer New York:, 2006.
- [4] M. Dahlbom and E.J. Hoffman. An evaluation of a two-dimensional array detector for high resolution PET. *Medical Imaging, IEEE Transactions on*, 7(4):264–272, 1988. ISSN 0278-0062.
- [5] B.J. Frey and D. Dueck. Clustering by passing messages between data points. *science*, 315(5814):972, 2007.
- [6] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK, 1988.
- [7] Christof KnoB. *Evaluation and Optimization of the High Resolution Research Tomograph (HRRT)*. PhD thesis, 2004. URL <http://d-nb.info/972775056>.
- [8] D. Koller and N. Friedman. *Probabilistic graphical models: Principles and techniques*. The MIT Press, 2009.
- [9] M. Opper and D. Saad. *Advanced mean field methods: Theory and practice*. Massachusetts Institute of Technology Press (MIT Press), 2001.
- [10] G. Parisi and R. Shankar. Statistical field theory. *Physics Today*, 41:110, 1988.

- [11] J.G. Rogers, R. Nutt, M. Andreaco, and CW Williams. Testing 144-and 256-crystal BGO block detectors. *Nuclear Science, IEEE Transactions on*, 41(4):1423–1429, 1994. ISSN 0018-9499.
- [12] V. Sossi, H.W.A.M. de Jong, W.C. Barker, P. Bloomfield, Z. Burbar, M.L. Camborde, C. Comtat, L.A. Eriksson, S. Houle, D. Keator, et al. The second generation hrst-a multi-centre scanner performance investigation. In *Nuclear Science Symposium Conference Record, 2005 IEEE*, volume 4, pages 2195–2199. IEEE, 2005.
- [13] K.A. Stonger and M.T. Johnson. Optimal calibration of PET Crystal position maps using Gaussian mixture models. *Nuclear Science, IEEE Transactions on*, 51(1):85–90, 2004. ISSN 0018-9499.
- [14] M.J. Wainwright and M.I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.
- [15] K. Wienhard, M. Schmand, ME Casey, K. Baker, J. Bao, L. Eriksson, WF Jones, C. Knoess, M. Lenox, M. Lercher, et al. The ECAT HRRT: performance and first clinical application of the new high resolution research tomograph. *IEEE Transactions on Nuclear Science*, 49(1):104–110, 2002.
- [16] K. Xiaowen, L. Yaqiang, W. Shi, S. Xishan, and Z. Rong. A fast accuracy crystal identification method based on Fuzzy C-Means (FCM) Clustering Algorithm for MicroPET. In *International Conference on BioMedical Engineering and Informatics (BEMI) Record*, 2008.

Feed-forward Hierarchical Model of the Ventral Visual Stream Applied to Functional Brain Image Classification

3.1 Introduction

Significant progress has been made in the diagnostic decision-making processes and in predicting the onset and the course of brain disorders ([18]; [23]; [29]; [31]). The traditional endpoint diagnosis, clinical measurements and cognitive tests used in clinical trials have proved to be informative but have their own limitations in accurately quantifying the progression of brain disorders in an unbiased and objective manner ([4]; [21]). Advances in brain imaging technologies have enabled researchers to investigate and test novel biomarkers that could serve either as diagnostic tools to aid clinical decision-making or as surrogates, reflecting disease progression and underlying disease pathology (Biomarkers Definitions Working Group, 2001). Accordingly, there is a growing body of evidence in the literature showing

Published in Human Brain Mapping 35.1 (2014): 38-52.

that structural and functional brain imaging can be valuable tools for predicting and classifying gradually progressive neurological and psychiatric disorders such as Alzheimers disease (AD) ([10]; [19]; [25]; [28]; [33]). Although both PET and MRI imaging modalities have been found to be discriminative in various neurological disorders, there is disagreement in the community about which are most sensitive for particular disorders. Specifically, differences in sensitivity and specificity of structural Magnetic Resonance Imaging (MRI) and 2-deoxy-D-glucose (FDG) Positron Emission Tomography (PET) features in the prediction of early AD has been debated in the literature with no clear consensus ([7]; [25]). Nevertheless, AD research studies evaluating the diagnostic and predictive value of regional specific glucose metabolic rate and volume changes suggest the greater reliability of FDG PET over MRI in discriminating AD from subjects with intact and mild cognitive impairment ([7]; [20]; [25]). However, De Santi and Mosconi indicate image post-processing influences the outcome of discriminative analyses and subsequently, their predictive value.

Although advances in imaging have enabled researchers to visually inspect both functional and structural brain scans of disease, it is often difficult for the human observer to identify the subtle differences in the brain images that are often necessary for reliable disease classification. Furthermore, visual identification of brain diseases by a human observer is time consuming and error prone. Automated image analysis algorithms that can reliably discriminate the diseased from the healthy brain are preferred because they save time, are generally less prone to errors, are not influenced by rater bias or inter-rater differences in neuroanatomical expertise, and can identify subtle statistical correlations in the data. For preventative and longitudinal studies in large populations, automated image analysis is critically important to evaluate the data. To achieve automated and reliable image analysis and classification, we can use computer vision techniques that are designed to extract information from images.

Object recognition in images and video is an active area of research in the computer vi-

sion community. Finding objects is fundamentally related to pattern recognition where the presence of unique patterns of colors, edges, and/or textures are consistent with a particular class of object. Probabilistic models are particularly well suited for recognition problems because they provide a structured approach to modeling uncertainty and can be less sensitive to noise in the data. Object recognition systems often consist of a feature extraction component and a classifier. The feature extractor is used to identify properties of the objects that are most important in discriminating one object from another. The features along with a labeled training set are then used to train a classifier to map the features into a class label for each object the detection system is built to recognize. Although the overall process is simple, there are many subtleties in real world applications of detection systems such as object illumination, scale, occlusion, and orientation that affect accuracy. Most often we have a small set of images representing the objects to be recognized and do not have exhaustive examples at all possible scales, orientations, illuminations, etc. The challenge is therefore to find a feature space that avoids irrelevant variations in the objects and instead captures the most discriminating characteristics ([12]). One source of inspiration for engineering such invariant features is the primate visual system, which performs object detection robustly across a huge range of viewpoints, illuminations and occlusions. One very successful method, the Scale Invariant Feature Transform (SIFT) proposed by Lowe [24] uses features with partial invariance to local variations in scale and illumination, similar to the receptive fields of the neurons in the inferior temporal cortex, an area important for object recognition in primates. Serre et al. [32] introduced a filtering method whose hierarchical architecture was designed specifically to emulate visual processing in the cat and primate striate cortex. They applied this method to detecting objects in photographs and reported high success rates from a few training examples. Mutch and Lowe [27] reported similar performance results using a similar filtering scheme that scaled the input images instead of the filters as was done in the Serre work.

Similar to object recognition in photographs, for automated image-based diagnosis, it is necessary to ignore some classes of variation across healthy individuals while identifying other specific variations which are indicative of disease state. Differences in ligand uptake in the brain measured by functional brain imaging modalities such as FDG PET and Tc99m HMPAO Single Photon Emission Tomography (SPECT) result in spatially smooth patterns of differing intensities which can be used to differentiate a disease group from healthy subjects. Similarly, precise morphology/anatomy may vary among individuals requiring some degree of local scale and orientation invariance. Based on this insight, we extend the neurologically-inspired filtering model described by Serre et al. [32] to signal detection in functional brain imaging. To evaluate how well the Serre feature model works in capturing disease patterns in the human brain, the model is extended to 3D volumetric space and signal detection differentiation in functional brain imaging. The hierarchical filtering pipeline is analyzed to identify which steps are most important for classification accuracy and the filter outputs are used to train both neural network (NN) and logistic regression (LR) classifiers. Two distinct and previously published datasets are tested using this feature extraction and classification method: (1) Alzheimers Disease Neuroimaging Initiative (ADNI) AD FDG PET scans sampled at baseline, 12 month, and 24 month time-points versus the study specific age-matched healthy comparison (HC) subjects ([26]); (2) a Tc99m HMPAO SPECT National Football League (NFL) dataset versus study specific age-matched HC subjects ([1]). The AD classification results are further compared against a blinded expert human rater (co-author J.H. Fallon), providing a baseline measure of how well a human counterpart can recognize disease in the same dataset.

3.2 Methods

3.2.1 Filtering and Feature extraction

The image filtering pipeline consists of a series of alternating steps of simple filtering (S layers) and complex filtering (C layers) layers briefly summarized here and discussed in detail in subsequent sections. The first simple layer (S1) outputs respond to oriented edges at different spatial scales and orientations (section 3.2.1.1). Spatial scales in this context refer to the underlying spatial distribution of the signal in the images. Filters with larger spatial scales will respond to larger (spatially) image signals. S1 layer filters are separated into bands where each band is composed of two similar spatial scales as shown in table 3.1, rows 1 and 2. The first complex layer (C1) combines the outputs from the S1 layer at different scales but within orientations, providing scale invariance (section 3.2.1.2). The complex layers pool the simple layer outputs using a max operator, where the strongest simple layer output drives the complex layer output. The second simple layer (S2) matches the detections from the C1 layer against healthy subjects in a template matching framework where higher scores indicate a closer match (sections 3.2.1.2.1 and 3.2.1.3). The second complex layer (C2) combines the outputs from template matching scores across orientations gaining invariance to orientation (section 3.2.1.4).

3.2.1.1 S1 Layer

The S1 layer is computed by applying sixteen orientated 3D Gabor filters at orientations $\Theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$, $\phi \in \{0, \pi/4, \pi/2, 3\pi/4\}$, and wavelength λ to each brain scan in the dataset. A Gabor filter is a linear filter whose impulse response is a harmonic function

multiplied by a Gaussian function:

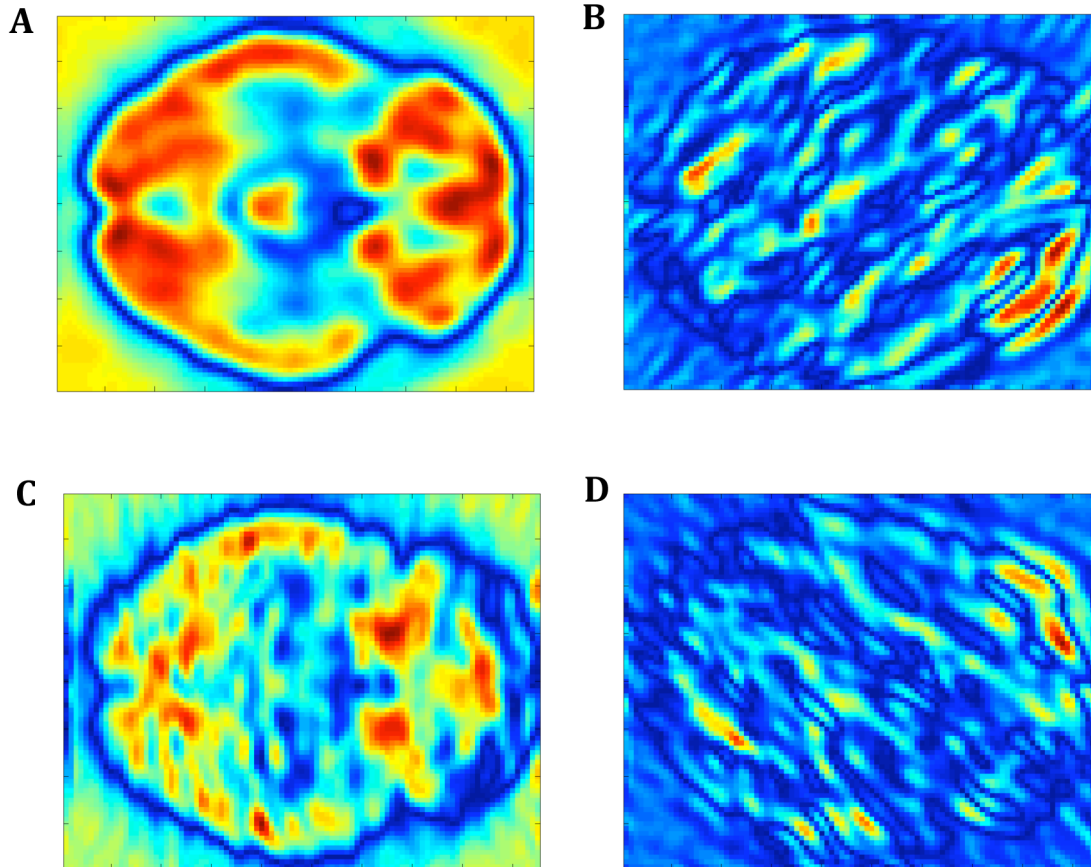
$$G(x, y, z) = \frac{1}{(2\pi)^{\frac{3}{2}}\sigma^3} \exp\left(-\frac{1}{2}\left(\frac{x^2+y^2+z^2}{\sigma^2}\right)\right) \cos\left(\frac{2\pi}{\lambda}(x \cos(\Theta) \sin(\phi) + y \sin(\Theta) \sin(\phi) + z \cos(\phi))\right) \quad (3.1)$$

The cosine term in equation 3.1 controls the harmonic component through the λ wavelength parameter. The variables x , y , and z are the spatial variables defining the spatial extent of the filter. The standard deviation σ describes the size of the Gaussian envelope. The orientation of the filter is represented by variables Θ and ϕ where ϕ orients the filter in the x-y plane and Θ is the orientation from the positive z axis. For a detailed description of 3D Gabor filters, refer to Bau et al. ([2]; [3]). Frequency and orientation representations of the filter are similar to those of the human visual system. The original Serre method performed Gabor filtering in 2D, consistent with the image matrix of photographs. In this work, the Gabor filtering was performed in 3D and applied using filter sizes, sigmas, and lambdas over a series of eight bands. The parameters of each band are listed in Table 3.1, rows 2-4. The filter sizes and parameters were kept essentially the same as the Serre work, but the spatial extents of the bands were decreased in order to make the features more sensitive to small activation differences in functional brain imaging. The relative proportions between sizes across the bands remained the same. The voxel sizes of the functional brain imaging data used in this study were 2mm^3 per voxel (see section Materials/Methods for a detailed description of the test data). The smallest filter size in the Serre work (7 pixels) if directly applied as 7 voxels would be unlikely to respond to small differential signals that could be discriminative in the context of functional imaging and disease. To avoid missing small signals, the lowest filter band was set to 3 voxels. An example of the AD PET scan slices filtered with the 3D Gabor functions are shown in Figure 3.1. Oriented signals are indeed differentially selected by the filters, consistent with our hypothesized responses of the filters when applied to functional brain imaging data.

Table 3.1: S1 Layer Gabor filter sizes and parameters by band (rows 1-3). Row 4 shows the C1 layer grid size for maximums over Gabor filter scales. Row 5 shows the template patch sizes common to all bands.


Band	1	2	3	4	5	6	7	8
Filter	3, 5	7, 9	11, 13	15, 17	19, 21	23, 25	27, 29	31, 33
Sigma	1.4, 2.1	3.0, 3.9	4.6, 5.6	6.5, 7.5	8.5, 9.6	10.6, 11.7	12.9, 14.1	15.3, 16.5
Lambda	1.7, 2.6	3.6, 4.8	5.7, 6.8	8.0, 9.2	10.4, 11.8	13.1, 14.5	15.9, 17.4	18.9, 20.5
Max Grid	4^3	6^3	8^3	10^3	12^3	14^3	16^3	18^3
Patch	$5^3, 9^3, 13^3, 17^3$							

Figure 3.1: Examples of Gabor filtered slices. For each example, the filter size, σ , and λ remained constant at 5^3 , 2.1, and 2.6 respectively while the orientation parameters Θ and ϕ were varied. A) $\Theta = 0$, $\phi = 0$; B) $\Theta = \pi/4$, $\phi = \pi/4$; C) $\Theta = \pi/2$, $\phi = \pi/4$; D) $\Theta = 3\pi/4$, $\phi = \pi/4$. The maximum filter responses are shown in red. As the orientation of the filters change (A-D), signals of similar orientations are selected by the filter.



3.2.1.2 C1 Layer

The C1 layer combines incident S1 units of the same Θ and ϕ orientations, creating tolerance to size and shift within Gabor filter orientation. Complex cells in the hierarchical visual cortex model have larger receptor fields than the S1 layer ([32]). To operationalize this relationship, the S1 layer volumes are filtered with a max operator over Gabor filter scales (Table I, row 1(filter)), but within each orientation band (columns of Table 3.1). Max filtering is a non-linear image processing technique where the value at each voxel in the filtered image is the maximum of the input image voxels in a local neighborhood defined by the filter size. The filter size over which the maximums are calculated depends on the Gabor filter size (shown in Table 3.1, row 4 (max grid)). Gabor filters with larger spatial scales will respond more strongly to larger (spatially) signals in the images at the same Θ and ϕ orientations, therefore, the corresponding max filter sizes should be tuned accordingly. These operations are performed for each Gabor orientation and for each band resulting in 16*8 volumes, representing maximums over scales but within orientations. Due to the large numbers of voxels in the volumes and thus the large numbers of max operations over increasing neighborhoods, we used the algorithm developed by Van Herk [36] to efficiently compute the maximums over neighborhoods for each voxel in the S1 layer volumes. The method requires only a small number of operations per voxel to compute the maximums and lowers the computational time of this stage of the processing pipeline.

 **3.2.1.2.1 C1 Layer Training Patches** Template matching is a common approach to object recognition in computer vision systems. It is a technique which matches image regions to stored representative templates using a specific scoring function ([6]). In this work, representative templates were collected on a random subset of hold-out healthy subjects to be used in the subsequent S2 layer template-matching step. Ten randomly selected hold-out training images were chosen for template extraction. Templates were extracted randomly across these training images and from random locations within the images but constrained

to fall within the boundaries of user specified regions of interest (see section 3.3.2). The regions from which templates are randomly sampled are completely user defined and could be chosen based on some a-priori hypothesis or from the literature. Selecting templates from specific regions of interest in the brain is similar to learning that a car is characterized by particular features in spatial locations, e.g. rides on four tires, has doors on the sides, a hood on the front, etc.

Operationally, the user selects regions of interest and the number of features prior to pipeline execution. We uniformly divide the number of random locations across the number of regions of interest. To generate the random voxel locations within a region of interest, we use an atlas labelmap, which assigns a numerical code to each atlas region. Each atlas region is therefore defined by all the contiguous voxels in the labelmap volume that have equal numerical codes. From this information, we can find the cube containing this region. We then use rejection sampling: drawing a random point uniformly within the cube, we accept it if it falls within the bounded region; otherwise we reject and try again. This process continues until the required number of locations has been found for each region. In our experiments, 50 or 100 templates were chosen to describe the low level representation of the brain images. We chose the two sets such that we had a reasonable number of templates per region of interest selected and so we could evaluate the dependence of the classification results on the number of feature scores used. The original Serre work suggests a modest dependence of performance on the number of feature scores used. For each selected template location, 5^3 , 9^3 , 13^3 , and 17^3 voxel patches were extracted from each of the 16 Gabor filtering orientations and bands from the C1 layer of the ten randomly selected hold-out healthy subject training images. These patches are simply contiguous sets of voxels of differing spatial extents (5^3 , 9^3 , 13^3 , and 17^3) centered on the template location and effectively give the vision system a memory of image feature examples from the functional brain images of healthy subjects.

3.2.1.3 S2 Layer

The S2 layer corresponds to the template-matching phase of the pipeline. For each C1 image in the test dataset and for each template patch collected from the hold-out healthy subject data, we compare the Gaussian radial basis function score shown in equation (3.2) for each band independently. The S2 units response depends on the Euclidean distance between the test dataset patch (X) and the stored prototype patch (P) sampled at the same location, scale, and orientation. If the functional activity profile in the test data is identical to the stored template patch, the score equals 1 whereas if the differences from the stored template patch are large, the score approaches 0. The parameter γ normalizes for different patch sizes ($n \in \{5, 9, 13, 17\}$) when computing the score in equation (3.2). The parameter γ is fixed to $(n/5)^3$ where n is the patch size and the denominator is the smallest patch size. The parameter σ in equation (3.2) is the uncertainty or variance in the stored prototype patch (P). This parameter was set to 1 in all experiments. Alternatively, it could be set to the empirical variance of the training prototype patches discussed in section 3.2.1.2.1.

$$F(X_{\Theta,\phi}, P_{\Theta,\phi}) = \exp\left(-\frac{\|X_{\Theta,\phi} - P_{\Theta,\phi}\|^2}{2\sigma^2\gamma}\right) \quad (3.2)$$

3.2.1.4 C2 Layer

The final layer in the pipeline computes the maximum response of the S2 layer scores from all bands and orientations for each prototype template. The final feature sets therefore consist of 50 or 100 shift and scale invariant scores (i.e., for 50 and 100 prototype patches) that are subsequently used for classification. Conceptually, for each test image, for each prototype

template patch sampled from a brain region, we are using the score that indicates the best match between the test image and a healthy subject regardless of signal size and orientation. We expect that subjects with neurological disorders will match less well with the healthy subjects and thus have a lower score. The final size of the feature vector therefore depends only on the number of patches extracted during training and not on the number of voxels in the full three-dimensional brain image. This allows the user to balance the number of template patches sampled during patch selection (i.e. number of features) and the number of subjects available in the dataset. Flexibility in choosing the number of features provides insulation from classifier over-fitting, which can occur if the number of features greatly exceeds the number of examples.

3.3 Evaluation

We used two datasets to evaluate the approach. Both are functional imaging datasets but distinctly different modalities. We selected these datasets to evaluate the generality of this approach and its application to distinctly different neurological abnormalities.

3.3.1 The Alzheimer's Disease Neuroimaging Initiative (ADNI)

ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations as a \$60 million, 5-year publicprivate partnership. The primary goal of ADNI is to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, such as cerebrospinal fluid (CSF) markers, APOE status and full-genome genotyping via

blood sample, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and AD. Determination of sensitive and specific markers of very early AD progression is intended to: (1) aid in the development of new treatments, (2) increase the ability to monitor their effectiveness, and (3) reduce the time and cost of clinical trials. The principal investigator of the initiative is Michael W. Weiner, M.D., of the Veterans Affairs Medical Center and University of California, San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and participants have been recruited from over 50 sites across the U.S. and Canada. ADNI participants range in age from 55 to 90 years and include approximately 200 cognitively normal elderly followed for three years, 400 elderly with MCI followed for three years, and 200 elderly with early AD followed for two years. Participants are evaluated at baseline, 6, 12, 18 (for MCI only), 24, and 36 months (AD participants do not have a 36 month evaluation). Baseline and longitudinal follow-up structural MRI scans are collected on the full sample and ^{11}C -labeled Pittsburgh Compound-B (^{11}C -PIB) and FDG PET scans are collected on a subset every 6 to 12 months (for study details see <http://www.adni-info.org>). A subset of these data were published in Mueller et al. [26] and Langbaum et al. [22] was used in this analysis.

3.3.2 AD Dataset

The dataset used in this study consisted of 154 baseline FDG PET scans acquired as part of the ADNI study and published in Mueller et al. [26] and Langbaum et al. [22]. There were 82 HC subjects (Mini-Mental State Exam (MMSE) 28.6 ± 1.1 ; Age 75.1 ± 9.6 yrs) and 72 AD subjects (MMSE 23.2 ± 3.5 ; Age 75.1 ± 11.2 yrs) from the baseline ADNI sample used for this study. The 12m and 24m ADNI samples contained a subset of the baseline dataset due to subject dropout. The 12m sample included 72 HC subjects (MMSE 29.2 ± 1.2 ; Age 77.5 ± 8.4 yrs) and 61 AD subjects (MMSE 20.9 ± 4.9 ; Age 75.4 ± 11.8 yrs). The 24m sam-

pled included 68 HC subjects (MMSE 28.6 ± 3.7 ; Age 76.0 ± 10.2 yrs) and 33 AD subjects (MMSE 18.4 ± 6.1 ; Age 74.6 ± 15.2 yrs). The acquisition protocol consisted of collecting six five-minute frames 30-60 minutes post 18FDG-injection. During the uptake period subjects were asked to rest comfortably in a dimly lit room with their eyes open. The collected frames were registered to the first frame (acquired at 30-35 min post-injection) and averaged to yield a single 30 minute average PET image in “native” space. The image matrix, field of view, and resolution of the datasets from participating sites were then matched by the ADNI group. The images were spatially normalized to the MNI atlas using SPM8 software (2007) resulting in image matrices of 79 x 95 x 68 voxels in x, y, and z dimensions respectively with isotropic 2³mm voxel sizes. The Automated Anatomical Labeling (AAL) atlas was used to constrain the region of interest selection based on the anatomical parcellations available in the atlas [35]. The AAL atlas used to define the region of interest boundaries is consistent with the space defined by the MNI atlas.

Coordinates for template patch sampling and S2 layer matching scores were constrained to fall within regions identified in the literature to be affected by AD (see sections 3.2.1.2.1 and 3.2.1.3). Delacourte et al. [8] identified stages of AD neurofibrillary degeneration in patients of various ages and different cognitive statuses. Further, Langbaum et al. [22] identified regions of reduced metabolic rates in AD. Regions included the cingulate cortex, parietal and temporal lobes, among others. For this study, we chose AAL atlas regions (bilateral): anterior and posterior cingulum, temporal lobes (middle), hippocampus, amygdala, thalamus, frontal and orbital cortices (superior and middle), temporal pole (superior, middle, inferior), and parietal lobe (inferior) as being consistent with published findings on potentially discriminative regions.

3.3.3 NFL Dataset

The NFL dataset used in this study consisted of 162 technetium-99m hexamethylpropyleneamine oxide (Tc99m HMPAO) SPECT scans acquired for a study evaluating the impact of playing American football by Amen et al. [1]. There were 83 HC (Age 41.7 ± 17.8 yrs) and 79 NFL (Age: 57.5 ± 11.5 yrs) subjects. Subjects were injected with an age/weight appropriate dose of Tc99m HMPAO and performed the Connors Continuous Performance test II for 30 minutes during uptake. All subjects completed the task and were subsequently scanned on a high-resolution Picker Prism 3000 triple-headed gamma camera with fan beam collimators. The original reconstructed image matrices were 128x128x29 voxels with sizes of 2.16mm x 2.16mm x 6.48mm. The images were spatially normalized to the MNI atlas using SPM8 software [13] resulting in image matrices of 79 x 95 x 68 voxels in x, y, and z dimensions respectively with isotropic 23mm voxel sizes. Images were smoothed using an 8mm FWHM isotropic Gaussian kernel. The pre-processing steps were identical to the previously published work by Amen et al. [1]. In the previously published work, a subset of the HC dataset was used and matched on gender and race. For this work, all subjects were used regardless of race and gender.

Coordinates for template patch sampling and S2 layer matching scores (see sections 3.2.1.2.1 and 3.2.1.3) were constrained to fall within regions identified in Amen et al. [1] as the top discriminating regions for the NFL group. To our knowledge, the Amen study was the first brain imaging study evaluating NFL players and as such, the regions were picked based only on that publication. For this study, we used AAL atlas regions (bilateral): anterior and posterior cingulum, frontal pole, hippocampus, amygdala, and temporal pole (middle and inferior).

3.3.4 Ethics

The NFL study was approved by each of the participating sites Institutional Review Boards (IRBs) and complied with the Code of Ethics of the World Medical Association (Declaration of Helsinki). Written informed consent was obtained from all participants after they had received a complete description of the studies. The ADNI data was previously collected across 50 research sites. Study subjects gave written informed consent at the time of enrollment for imaging and genetic sample collection and completed questionnaires approved by each participating sites Institutional Review Board (IRB).

3.3.5 Feature Sets

In order to identify which components of the feed-forward hierarchical model implemented in this study were most important in correct classification, three separate feature sets were computed. The FTM (Gabor filter + template match) feature set is the result of the full hierarchical pipeline as described in section 4.2. In order to understand the effect of the Gabor filtering, the TM (template matching) dataset was created using the same procedures outlined in section 4.2 without Gabor filtering. More precisely, the dataset consists of selecting template patches from the un-filtered images (neither S1 nor C1 layers) and performing the computations in the S2 and C2 layers. To evaluate the effect of template matching, the AP (average patch) feature set consists of simply averaging the voxels in the neighborhood around the prototype patch locations selected in section 3.2.1.2.1, across the various filter sizes (Table 3.1, row 2) and taking the maximum response.

To compare the feature sets of the hierarchical model with more typical data reduction techniques, the maximum group difference (MaxT) and data reduction (DR) sets were com-

puted. The MaxT feature set is computed by performing a typical voxel-wise independent, 2 sample t-test in the SPM8 software. The resulting SPM(t) maps were then thresholded at $p < 0.01$ (AD baseline), $p < 0.001$ (AD 12m), $p < 0.001$ (AD 24m), and $p < 1e - 6$ (NFL) and corrected for multiple comparisons using the family-wise error rate (FWE) correction. Probability thresholds were chosen to limit the number of voxels in the resulting t-score maps such that similar numbers of voxels were obtained for each data set (3K points). The absolute values of the resulting t-scores were ranked and the data from the top 50 and 100 locations were then sampled from each subject and used for classification (MaxT). The DR feature set used all the locations found in the group difference maps, discussed above, after probability thresholding (3K points), sampling the original data at those locations (3K points) for each subject. The resulting NxK matrix (N subjects, K sampling locations) was mean-centered for each column K and run through principal components analysis (PCA). Each subjects data was then projected onto the eigenvectors of the top 50 and 100 largest eigenvalues from the PCA decomposition giving a low dimensional representation with 50 or 100 feature scores that were subsequently used for classification. The top 50 and 100 largest eigenvectors were chosen so that the projected dataset contained 50 and 100 scores per subject, consistent with the number of feature scores calculated from the full feed-forward hierarchical model.

3.3.6 Classification

Classification was done using both a multilayered perceptron neural network (NN) and a logistic regression (LR) classifier to understand the dependence of the results on the classifier chosen [15]. Each classifier was trained separately on the same datasets to compare the performance of the simpler logistic regression classifier, able to find linear decision boundaries,

with the neural network classifier, able to model more complex nonlinear functions. The neural network was constructed with one hidden layer (hidden layer nodes = $(\#features + \#classes)/2$) and trained with a learning rate of 0.3 and a momentum of 0.2. Adding additional hidden layers may result in a marginal improvement in classification accuracy but at the expense of learning complexity. For each classifier, ten-fold cross validation was used. The dataset was divided in each fold into training and testing subsets. The classifier was trained using the training subset and tested on the testing subset. This process was repeated ten times. Areas under the receiver operating characteristic (ROC-AUC) curves were computed from the probability of class membership of the testing data from each of the trained classifiers. The full filtering pipeline ROC-AUC curves were statistically compared to each of the alternative methods for each dataset and classifier using the DeLong et al. [9] method of comparing areas under correlated ROC curves as implemented in the pROC package [30]. To compute 95% confidence intervals and statistics, the data was resampled 2000 times, stratified by group membership.

To compare the classifier results on the baseline AD dataset with the visual ratings of neuroanatomist and co-author J.H. Fallon, true positive (TP) and false positive (FP) rates were calculated. To calculate the TP and FP rates, the probability of class membership from the trained classifiers for each testing subset data point, in each fold, was computed. The data point was assigned to the class with the largest probability. The TP rate was the proportion of examples in the testing subsets that were classified as class AD, among all testing examples that were originally labeled as class AD. The TP rate is the average across all folds. The FP rate was the proportion of examples in the testing subsets that were classified as class AD, but were originally labeled as the alternative class, among all testing examples which are not of class AD. The FP rate is the average across all folds. The TP and FP results for Dr. Fallon were computed from his designation of either AD or healthy control for each of the baseline data compared to the original class labels.

3.4 Results

To summarize the performance of each classifier, the ROC-AUC results for the Alzheimers disease (AD) baseline, 12m, and 24m datasets are shown in Figures 3.2, 3.3, and 3.4 respectively for 50 and 100 feature datasets and both logistic regression and neural network classifiers. The confidence intervals for each ROC-AUC and statistical comparisons of the full filtering pipeline (FTM) with each of the other methods for all classifiers and datasets are shown in Tables 3.2, 3.3, and 3.4. The FTM method outperformed the other methods in terms of ROC-AUC in 80% of the tests, and was statistically better in 35%. No other method was statistically better than FTM; although, the PCA data reduction strategy (DR) in the 50-feature, baseline AD, logistic regression classifier was close ($p < 0.064$). Overall, the neural network classifier generally outperformed the logistic regression classifier in ROC-AUC. Further, the FTM method was statistically better than all other methods in 46% of the neural network classification experiments compared to 25% using the logistic regression classifier, suggesting a benefit of using the more sophisticated classifier with the FTM method. There was a small, non-significant, increase on average in ROC-AUC over all the classifiers in the results using the larger 100 feature datasets. Overall performance of the FTM trained classifiers were consistent with other published classification results (see Discussion) using the ADNI dataset, with maximum ROC-AUC at baseline of 0.962 ± 0.025 (neural network, 100 feature), at 12m of 0.837 ± 0.073 (neural network, 100 feature), and at 24m of 0.878 ± 0.070 (neural network, 100 feature).

Neuroanatomist and co-author J.H. Fallon was given the baseline AD dataset images in transaxial, coronal, and sagittal orientations, without the diagnosis and given no practice set of normal or ADs to examine prior to the analysis, and asked to classify the scans as either AD or HC. These results are only available for the baseline AD data due to the significant effort in manually rating so many scans. Dr. Fallon achieved a true/false positive rate for

Table 3.2: Results from the AD ROC-AUC analysis of the ADNI baseline data. The table lists ROC-AUC measurements, 95% confidence intervals, Z-scores, and probabilities for comparisons of the FTM method with the other methods within each dataset and classifier combination. Negative z-scores indicate methods that are lower in ROC-AUC than the FTM method. Significant differences are highlighted in bold (MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = gabor filtering + template matching).

Dataset (# feat)	Classifier	Method	ROC-AUC	95% Conf	Z-score ($X_{AUC} - FTM_{AUC}$)	P($FTM_{AUC} = X_{AUC}$)
AD – Bas (50)	LR	FTM	0.791	0.857-0.725		
		TM	0.644	0.721-0.567	-3.259	0.001
		AP	0.729	0.801-0.657	-1.556	0.120
		DR	0.861	0.919-0.803	1.854	0.064
		MaxT	0.692	0.768-0.616	-2.187	0.029
		AD-Bas (50)	NN	FTM	0.928	0.970-0.886
		TM	0.783	0.858-0.709	-4.321	1.55E-04
		AP	0.902	0.951-0.854	-1.132	0.258
		DR	0.905	0.952-0.858	-0.833	0.405
		MaxT	0.777	0.855-0.698	-3.661	2.51E-04
AD – Bas (100)	LR	FTM	0.763	0.832-0.694		
		TM	0.689	0.766-0.612	-1.761	0.078
		AP	0.713	0.832-0.694	-1.320	0.187
		DR	0.698	0.775-0.620	-1.604	0.109
		MaxT	0.687	0.761-0.614	-1.574	0.115
		AD-Bas (100)	NN	FTM	0.962	0.987-0.938
TM	0.644			0.722-0.567	-8.623	2.20E-16
AP	0.885			0.940-0.831	-3.336	8.50E-04
DR	0.678			0.763-0.594	-6.697	2.13E-11
MaxT	0.773			0.849-0.696	-5.053	4.35E-07

Table 3.3: Results from the AD ROC-AUC analysis of the ADNI 12m data. The table lists ROC-AUC measurements, 95% confidence intervals, Z-scores, and probabilities for comparisons of the FTM method with the other methods within each dataset and classifier combination. Negative z-scores indicate methods that are lower in ROC-AUC than the FTM method. Significant differences are highlighted in bold (MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = gabor filtering + template matching).

Dataset (# feat)	Classifier	Method	ROC-AUC	95% Conf	Z-score (X _{AUC} -FTM _{AUC})	P(FTM _{AUC} = X _{AUC})
AD-12m (50)	LR	FTM	0.778	0.851-0.705		
		TM	0.664	0.747-0.582	-2.173	0.030
		AP	0.756	0.831-0.682	-0.499	0.618
		DR	0.726	0.805-0.648	-1.060	0.289
		MaxT	0.609	0.701-0.518	-2.830	0.005
AD-12m (50)	NN	FTM	0.825	0.898-0.753		
		TM	0.781	0.862-0.701	-0.952	0.341
		AP	0.838	0.908-0.769	0.319	0.750
		DR	0.771	0.854-0.689	-1.292	0.196
		MaxT	0.681	0.776-0.585	-2.371	0.019
AD-12m (100)	LR	FTM	0.759	0.835-0.683		
		TM	0.648	0.734-0.561	-2.166	0.030
		AP	0.699	0.781-0.618	-1.210	0.226
		DR	0.676	0.763-0.588	-1.546	0.122
		MaxT	0.671	0.754-0.588	-1.532	0.127
AD-12m (100)	NN	FTM	0.837	0.910-0.764		
		TM	0.783	0.861-0.706	-1.411	0.158
		AP	0.855	0.919-0.791	0.590	0.555
		DR	0.714	0.802-0.627	-2.234	0.022
		MaxT	0.687	0.780-0.594	-2.482	0.014

Table 3.4: Results from the AD ROC-AUC analysis of the ADNI 24m data. The table lists ROC-AUC measurements, 95% confidence intervals, Z-scores, and probabilities for comparisons of the FTM method with the other methods within each dataset and classifier combination. Negative z-scores indicate methods that are lower in ROC-AUC than the FTM method. Significant differences are highlighted in bold (MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = gabor filtering + template matching).

Dataset (# feat)	Classifier	Method	ROC-AUC	95% Conf	Z-score (X _{AUC} -FTM _{AUC})	P(FTM _{AUC} = X _{AUC})
AD-24m (50)	LR	FTM	0.749	0.843-0.655		
		TM	0.658	0.763-0.553	-1.437	0.151
		AP	0.794	0.886-0.702	0.794	0.427
		DR	0.828	0.915-0.740	1.371	0.171
		MaxT	0.787	0.902-0.673	0.502	0.616
		FTM	0.841	0.924-0.758		
AD-24m (50)	NN	TM	0.883	0.955-0.810	0.991	0.322
		AP	0.865	0.942-0.788	0.736	0.462
		DR	0.816	0.915-0.717	-0.459	0.646
		MaxT	0.766	0.861-0.670	-1.335	0.182
		FTM	0.822	0.906-0.737		
		TM	0.823	0.906-0.740	0.026	0.979
AD-24m (100)	LR	AP	0.822	0.892-0.716	-0.319	0.75
		DR	0.561	0.426-0.696	-4.620	3.84E-06
		MaxT	0.813	0.915-0.710	-0.143	0.886
		FTM	0.878	0.948-0.806		
		TM	0.864	0.944-0.783	-0.383	0.702
		AP	0.880	0.957-0.804	0.102	0.919
AD-24m (100)	NN	DR	0.677	0.788-0.566	-8.214	2.20E-16
		MaxT	0.758	0.860-0.656	-2.273	0.023
		FTM	0.878	0.948-0.806		
		TM	0.864	0.944-0.783	-0.383	0.702
		AP	0.880	0.957-0.804	0.102	0.919
		DR	0.677	0.788-0.566	-8.214	2.20E-16

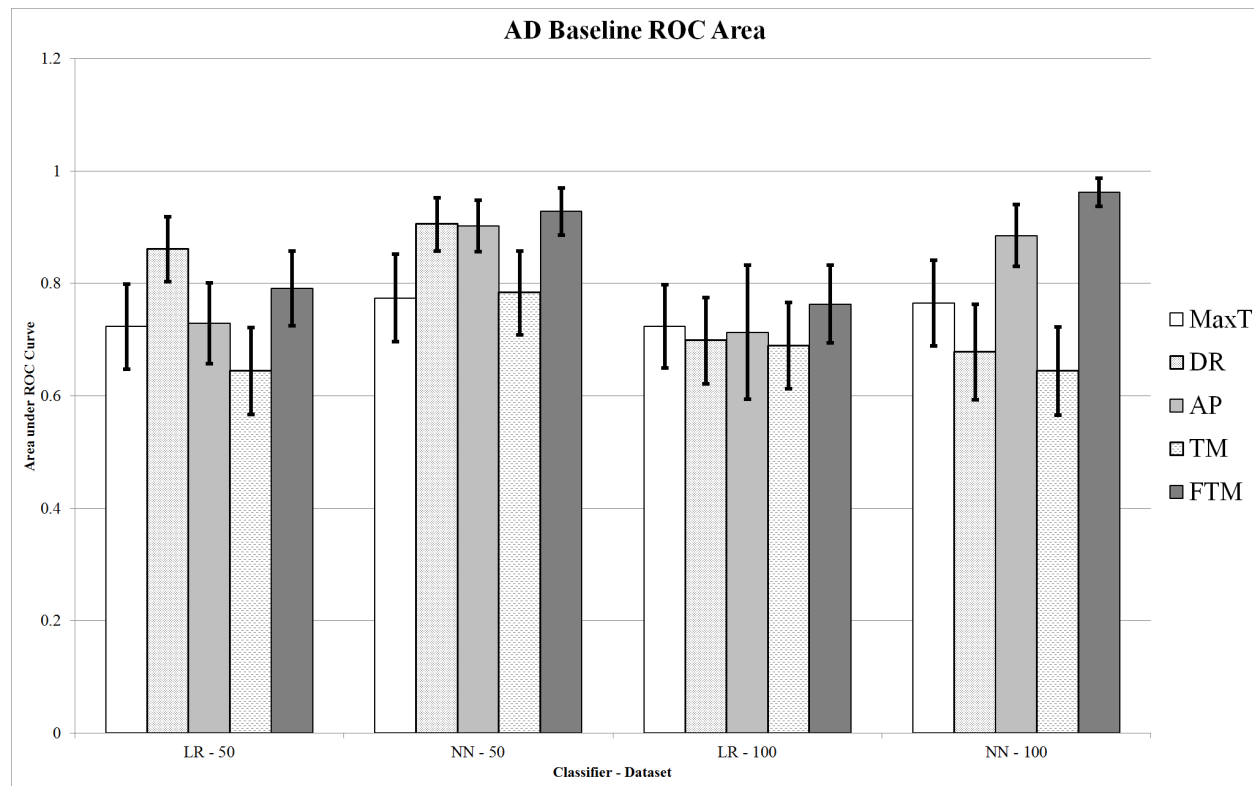
Table 3.5: Results from the NFL ROC-AUC analysis. The table lists ROC-AUC measurements, 95% confidence intervals, Z-scores, and probabilities for comparisons of the FTM method with the other methods within each dataset and classifier combination. Negative z-scores indicate methods that are lower in ROC-AUC than the FTM method. Significant differences are highlighted in bold (MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = gabor filtering + template matching).

Dataset (# feat)	Classifier	Method	ROC-AUC	95% Conf	Z-score ($X_{AUC} - FTM_{AUC}$)	P($FTM_{AUC} = X_{AUC}$)
NFL (50)	LR					
		FTM	0.909	0.954-0.864		
		TM	0.702	0.777-0.628	-4.662	5.09E-06
		AP	0.876	0.931-0.821	-0.907	0.365
		DR	0.942	0.979-0.906	1.137	0.255
		MaxT	0.956	0.988-0.924	2.059	0.040
NFL (50)	NN					
		FTM	0.908	0.957-0.860		
		TM	0.856	0.917-0.795	-1.306	0.193
		AP	0.933	0.974-0.893	0.776	0.438
		DR	0.974	0.995-0.954	2.405	0.016
		MaxT	0.986	1.000-0.969	2.944	0.003
NFL (100)	LR					
		FTM	0.939	0.976-0.902		
		TM	0.877	0.931-0.822	-1.856	0.065
		AP	0.883	0.936-0.830	-1.705	0.089
		DR	0.900	0.947-0.853	-1.315	0.188
		MaxT	0.977	1.000-0.954	1.753	0.080
NFL (100)	NN					
		FTM	0.920	0.964-0.876		
		TM	0.954	0.989-0.918	1.167	0.245
		AP	0.942	0.977-0.907	0.765	0.445
		DR	0.892	0.941-0.842	-0.866	0.386
		MaxT	0.988	1.000-0.973	2.900	0.004

Table 3.6: Results from the visual ratings of neuroanatomist J.H. Fallon from the ADNI baseline data. The table lists true positive (TP) and false positive (FP) values for the Alzheimers disease (AD) and healthy control (HC) classes compared to the FTM, DR, and MaxT methods. The FTM method outperforms both the human rater and the other methods (MaxT = maximum t-score, DR = PCA data reduction, FTM = gabor filtering + template matching).

Method	AD-TP	AD-FP	HC-TP	HC-FP
J.H. Fallon	0.718	0.380	0.671	0.244
FTM	0.875	0.122	0.878	0.125
DR	0.622	0.389	0.611	0.378
MaxT	0.829	0.375	0.625	0.171

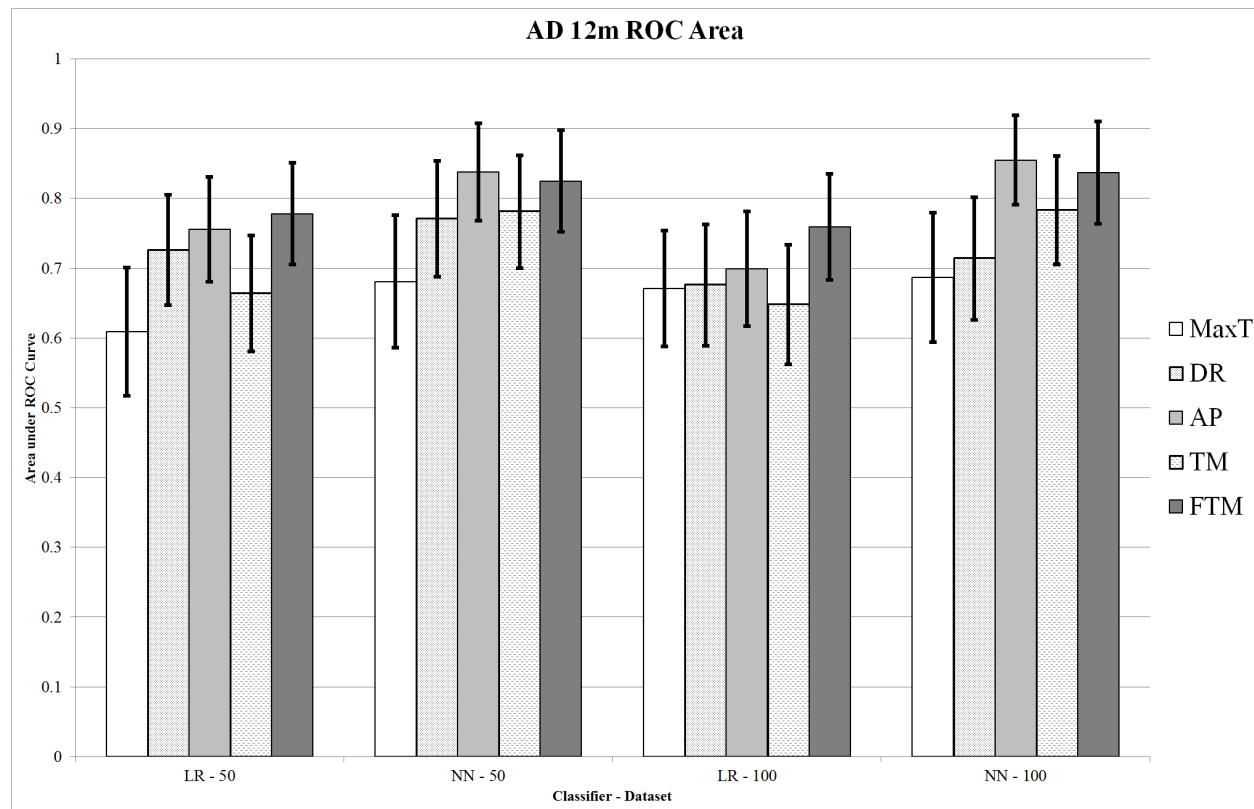
Figure 3.2: Area under the ROC curve for AD classification of the ADNI baseline data set for logistic regression (LR) and neural network (NN) classifiers for both 50 and 100 feature datasets (MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = gabor filtering + template matching). The FTM method outperforms the others in 94% of the cases and is statistically better in 50% of the cases.



AD of 0.718/0.380 and for the HC group of 0.671/0.244 as shown in Table 3.6. The FTM classifier performed better in both true/false positives for both AD and HC groups while also outperforming the maximum group difference (MaxT) and data reduction (DR) methods, further suggesting the potential utility of this approach.

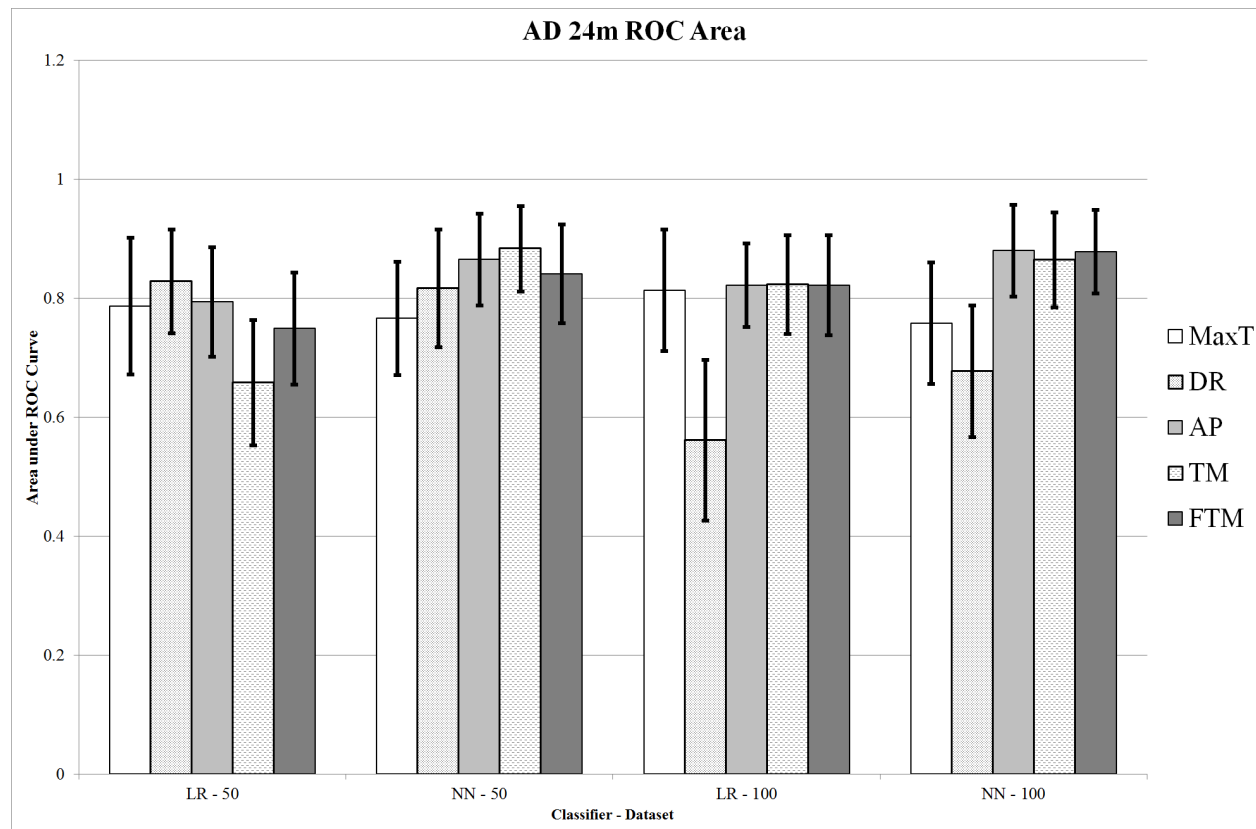
The AUC results for the NFL group are shown in Figure 3.5 for 50 and 100 feature datasets and both logistic regression and neural network classifiers. The confidence intervals for each ROC-AUC and statistical comparisons of the FTM with each of the other methods for all classifiers and datasets are shown in Table 3.5. Interestingly, unlike the AD dataset, the FTM method did not dominate the others, outperforming the other methods in 44% of the tests and was statistically better in only one. Alternatively, the MaxT method consistently

Figure 3.3: Area under the ROC curve for AD classification of the ADNI 12m data set for logistic regression (LR) and neural network (NN) classifiers for both 50 and 100 feature datasets (MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = gabor filtering + template matching). The FTM method outperforms the others in 88% of the cases and is statistically better in 38% of the cases.



outperformed the others in terms of ROC-AUC and was statistically better than the FTM method in three out of four comparisons. We speculate this result is related to specific brain functional changes accompanying repeated head injuries evident in the NFL dataset (see section 4.4). Overall performance of the FTM classifier was still quite good with maximum ROC-AUC of $0.939 \pm 0.037/0.145$ (logistic regression, 100 features). Unlike the AD experiments, the neural network classifier did not outperform the logistic regression classifier for the FTM dataset but did for the best performing MaxT dataset.

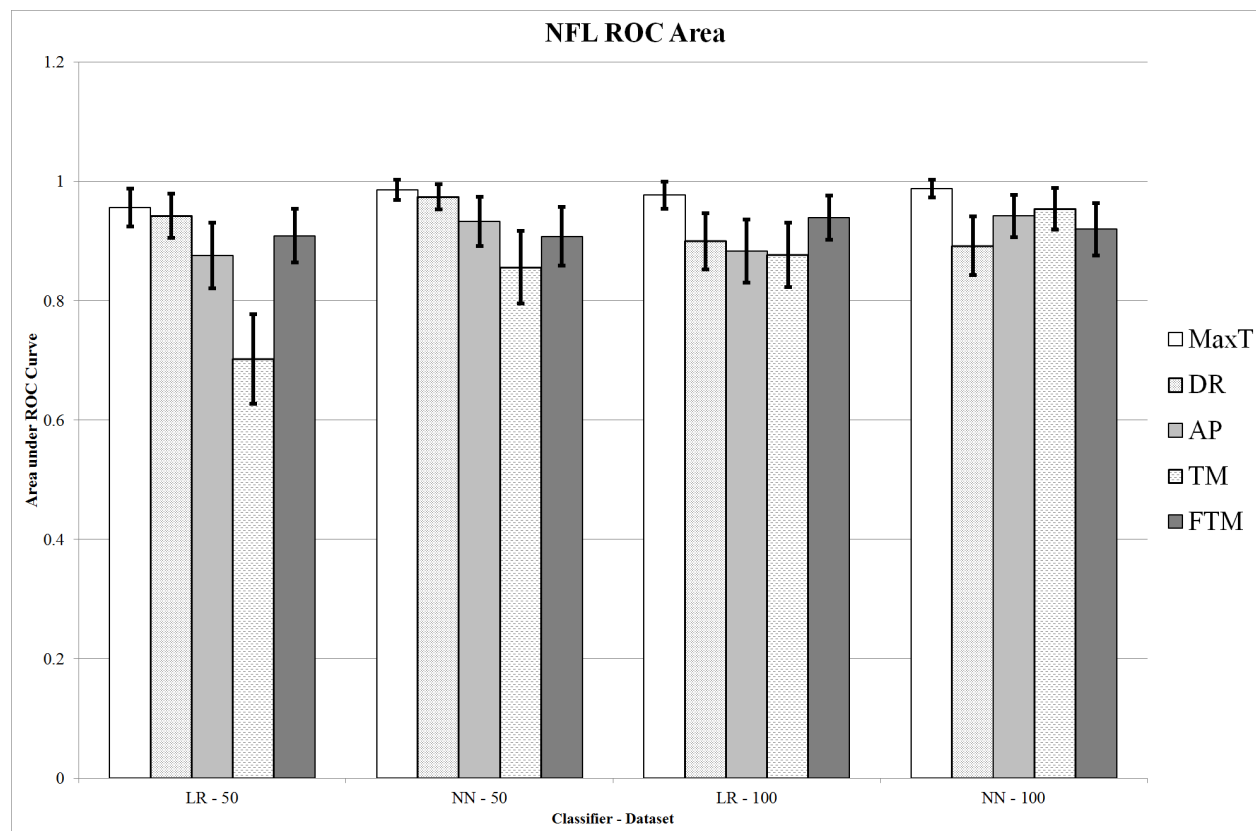
Figure 3.4: Area under the ROC curve for AD classification of the ADNI 24m data set for logistic regression (LR) and neural network (NN) classifiers for both 50 and 100 feature datasets (MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = gabor filtering + template matching). The FTM method outperforms the others in 56% of the cases and is statistically better in 19% of the cases.



3.5 Discussion

The overall classification results suggest the biophysically inspired feed-forward hierarchical model used in these experiments is sensitive to differences in functional brain imaging data. Both AD and NFL classification experiments showed impressive ROC-AUC rates using a method not specifically tuned for these imaging modalities. The full filtering pipeline (FTM) results are consistent with published classification rates for the ADNI AD data set using brain imaging; although, most reported results use a mix of structural and functional imaging features. For example, Hinrichs et al. [16] used the ADNI dataset in a spatially augmented boosting framework and reported an ROC-AUC of 0.8716 when using just FDG PET.

Figure 3.5: Area under the ROC curve for NFL classification for logistic regression (LR) and neural network (NN) classifiers for both 50 and 100 feature datasets (MaxT = maximum t-score, DR = PCA data reduction, AP = average patch, TM = template matching, FTM = gabor filtering + template matching). The MaxT method outperformed the other methods, statistically better than the FTM method in all comparisons except in the LR-50 feature dataset. The FTM ROC-AUC was still very good, always greater than 0.900 and as high as 0.939 in the NN-100 feature dataset.



A benefit of using logistic regression classifiers is the clear interpretation of which features are most informative for classification. For baseline AD classification, the four most informative patches (highest weights) were sampled from AAL atlas regions right hippocampus and superior temporal lobes left and right while the posterior cingulate, a region commonly associated with disease progression, ranked fourth. For 12m AD classification the most informative patches were sampled from frontal superior right, frontal superior orbital left, and the temporal pole superior right. For 24m AD classifications the most informative patches

were sampled from the frontal superior right, temporal pole mid left, and hippocampus left. It is interesting that the frontal lobe was not included as a top discriminating location in the baseline data set but was in both the 12m and 24m data, consistent with well-known structural changes in AD disease progression. We also evaluated the performance of the FTM features using ROIs that specifically did not include those selected in section 3.3.2. The results were on average 10-15% lower in ROC-AUC for baseline AD than those reported in the results section, suggesting this method is sensitive to region of interest selection. Therefore we suspect the filtering pipeline could be used to test competing hypotheses about specific regions of interest implicated in disease. The top three most informative patches from the features evaluated using ROIs that specifically did not include those selected in section 3.3.2, were sampled from AAL atlas regions frontal inferior orbital left, insula right, and occipital middle right. Other informative patches for AD included the supramarginal right, lingual right and frontal inferior operculum left. Interestingly, the frontal inferior orbital, operculum, and the supramarginal gyrus are all associated with AD in the literature suggesting the classification results are still picking up on areas related to the disease [11] [14].

Overall, the average patch (AP) feature set outperformed the template matching (TM) feature set, suggesting no compelling benefit of template matching without Gabor filtering in this application. The utility of oriented Gabor filtering and template matching in deriving the feature set was most evident in AD classification. This trend was not observed for the NFL classification experiments. Why would oriented filtering improve classification rates in AD and not the NFL data set? It is well known in the literature that structural changes in AD follow an anatomical trajectory starting in entorhinal cortex and hippocampus, then moving to temporal and parietal lobes, and finally affecting the frontal lobes in late stage AD [5] [17] [34]. These structural changes should be reflected in corresponding functional changes. In addition, the accumulation of amyloid plaques between nerve cells in the brain is known to be a hallmark of AD. Both the structural changes and plaques may be altering

the functional brain imaging derived signal in orientation, scale, and localized spatial extent due, in part, to brain plasticity and compensation.

Alternatively, the full filtering pipeline might not perform as well in data sets with widespread, global functional changes observed in the NFL data. Indeed the manuscript by Amen et al. reports significant decreases in regional cerebral blood flow were seen across the whole brain. The comparison feature sets MaxT and DR should perform well in that setting because they rely on group differences and maximal variation. It is possible that the FTM method performs better in settings with more localized functional differences. The NFL dataset differed from the AD dataset in both imaging modality (SPECT vs. PET) and uptake conditions (continuous performance test vs. rest), which could contribute to the differences in classifier performance. We suspect modality is not a factor as the feature scores used in classification are modality neutral. Lower resolution imaging systems may contribute to lower true positive rates if the regions of interest are small in size, despite the models attempt to mitigate this effect using filter sizes of differing spatial scales. Regardless of how well the filtering method does, if the discriminating feature of a disease is too small to be accurately measured by the imaging device, performance of the classification system will undoubtedly suffer. The benefit of this method is that it uses information across spatial scales, orientations, and locations in the volumes to calculate the matching scores used for subsequent classification and should therefore be less reliant on any one discriminating feature. The uptake task will contribute to the functional signals and should be taken into account when selecting the regions of interest to calculate feature scores (section 3.2.1.2.1). Choosing regions that are absolutely not affected by the disease will decrease the discriminative power of the method. Alternatively, if the number of subjects in the dataset is high and there is no fear of classifier overfitting, choosing many regions, some known to be related to the disease and/or task and others whose relationship is unknown could provide interesting insight into whether the unknown regions are contributing to classification accuracy. Further, because the features

of the dataset are computed separately from the classifier, one could choose to sample some features from all brain regions and either perform regularization in the classifier or choose a classification model that is less sensitive to overfitting (e.g. support vector machines). Each of these decisions should be made relative to the particular dataset and illness being studied.

3.6 Conclusions

In general our volumetric variant of the hierarchical feed-forward model originally proposed by Serre et al. [32] for detecting objects in photographs performed quite well on the functional brain imaging data sets used in this study. In fact, it outperformed both the comparison methods and the human counterpart at detecting AD in the FDG PET ADNI data set. The method is very general and does not rely on particular imaging modalities. It could be used on many spatial maps commonly computed in diagnostic and research imaging studies. Furthermore, there is evidence that it could be used to test hypotheses about regions implicated in disease. In conclusion, models designed in the computer vision community for object recognition and tracking in images of natural scenes may indeed have applications in detecting and tracking disease progression in human functional brain imaging with minimal modifications.

Bibliography

- [1] D. Amen, A. Newberg, R. Thatcher, Y. Jin, J. Wu, B. Phillips, D. Keator, and K. Willeumier. Impact of playing professional american football on long-term brain function. *Journal of Neuropharmacology*, 2010.
- [2] Tien C Bau and Glenn Healey. Rotation and scale invariant hyperspectral classification using 3d gabor filters. In *SPIE Defense, Security, and Sensing*, pages 73340B–73340B. International Society for Optics and Photonics, 2009.
- [3] Tien C Bau, Subhadip Sarkar, and Glenn Healey. Using three-dimensional spectral/s-patial gabor filters for hyperspectral region classification. In *SPIE Defense and Security Symposium*, pages 69660E–69660E. International Society for Optics and Photonics, 2008.
- [4] B. Borroni, E. Premi, M. Di Luca, and A. Padovani. Combined biomarkers for early Alzheimer disease diagnosis. *Current medicinal chemistry*, 14(11):1171–1178, 2007. ISSN 0929-8673.
- [5] H. Braak and E. Braak. Staging of Alzheimer-related cortical destruction. *International Psychogeriatrics*, 9(S1):257–261, 1997. ISSN 1041-6102.
- [6] R. Brunelli. *Template matching techniques in computer vision: Theory and practice*. John Wiley & Sons Inc, 2009. ISBN 0470517069.
- [7] S. De Santi, M.J. de Leon, H. Rusinek, A. Convit, C.Y. Tarshish, A. Roche, W.H. Tsui, E. Kandil, M. Boppana, K. Daisley, et al. Hippocampal formation glucose metabolism and volume losses in MCI and AD. *Neurobiology of aging*, 22(4):529–539, 2001. ISSN 0197-4580.

- [8] A. Delacourte, JP David, N. Sergeant, L. Buee, A. Wattez, P. Vermersch, F. Ghzali, C. Fallet-Bianco, F. Pasquier, F. Lebert, et al. The biochemical pathway of neurofibrillary degeneration in aging and Alzheimer's disease. *Neurology*, 52(6):1158, 1999.
- [9] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.
- [10] A. Drzezga. Diagnosis of Alzheimer's disease with [18F] PET in mild and asymptomatic stages. *Behavioural Neurology*, 21(1):101–115, 2009. ISSN 0953-4180.
- [11] Alberto J Espay and DH Jacobs. Frontal lobe syndromes, 2012.
- [12] D.A. Forsyth and J. Ponce. *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference, 2002. ISBN 0130851981.
- [13] K.J. Friston, J. Ashburner, S.J. Kiebel, T.E. Nichols, and W.D. Penny, editors. *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Academic Press, 2007. URL <http://books.elsevier.com/neuro/?isbn=9780123725608&srccode=89660>.
- [14] Y Grignon, C Duyckaerts, M Benneceb, and J-J Hauw. Cytoarchitectonic alterations in the supramarginal gyrus of late onset alzheimer's disease. *Acta neuropathologica*, 95(4):395–406, 1998.
- [15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009. ISSN 1931-0145.
- [16] C. Hinrichs, V. Singh, L. Mukherjee, G. Xu, M.K. Chung, and S.C. Johnson. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. *Neuroimage*, 48(1):138–149, 2009. ISSN 1053-8119.

- [17] X. Hua, A.D. Leow, N. Parikshak, S. Lee, M.C. Chiang, A.W. Toga, C.R. Jack Jr, M.W. Weiner, and P.M. Thompson. Tensor-based morphometry as a neuroimaging biomarker for Alzheimer’s disease: an MRI study of 676 AD, MCI, and normal subjects. *Neuroimage*, 43(3):458–469, 2008. ISSN 1053-8119.
- [18] K. Kantarci and C.R. Jack Jr. Neuroimaging in Alzheimer disease: an evidence-based review. *Neuroimaging Clinics of North America*, 13(2):197, 2003.
- [19] T. Kawachi, K. Ishii, S. Sakamoto, M. Sasaki, T. Mori, F. Yamashita, H. Matsuda, and E. Mori. Comparison of the diagnostic performance of FDG-PET and VBM-MRI in very mild Alzheimer’s disease. *European Journal of Nuclear Medicine and Molecular Imaging*, 33(7):801–809, 2006. ISSN 1619-7070.
- [20] T. Kawachi, K. Ishii, S. Sakamoto, M. Sasaki, T. Mori, F. Yamashita, H. Matsuda, and E. Mori. Comparison of the diagnostic performance of FDG-PET and VBM-MRI in very mild Alzheimer’s disease. *European Journal of Nuclear Medicine and Molecular Imaging*, 33(7):801–809, 2006. ISSN 1619-7070.
- [21] DS Knopman, S.T. DeKosky, JL Cummings, H. Chui, J. Corey-Bloom, N. Relkin, GW Small, B. Miller, and JC Stevens. Practice parameter: diagnosis of dementia (an evidence-based review): report of the Quality Standards Subcommittee of the American Academy of Neurology. *Neurology*, 56(9):1143, 2001.
- [22] J. Langbaum, K. Chen, W. Lee, C. Reschke, D. Bandy, A.S. Fleisher, G.E. Alexander, N.L. Foster, et al. Categorical and correlational analyses of baseline fluorodeoxyglucose positron emission tomography images from the Alzheimer’s Disease Neuroimaging Initiative (ADNI). *Neuroimage*, 45(4):1107–1116, 2009. ISSN 1053-8119.
- [23] S. Lovestone. Searching for biomarkers in neurodegeneration. *Nature Medicine*, 16(12):1371–1372, 2010. ISSN 1078-8956.

- [24] D.G. Lowe. Object recognition from local scale-invariant features. In *iccv*, page 1150. Published by the IEEE Computer Society, 1999.
- [25] L. Mosconi, S. Sorbi, M.J. de Leon, Y. Li, B. Nacmias, P.S. Myoung, W. Tsui, A. Ginestroni, V. Bessi, M. Fayyazz, et al. Hypometabolism exceeds atrophy in presymptomatic early-onset familial Alzheimer’s disease. *Journal of Nuclear Medicine*, 47(11):1778, 2006.
- [26] S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, C. Jack, W. Jagust, J.Q. Trojanowski, A.W. Toga, and L. Beckett. Alzheimer’s Disease Neuroimaging Initiative. *Advances in Alzheimer’s and Parkinson’s Disease*, pages 183–189, 2008.
- [27] J. Mutch and D.G. Lowe. Multiclass object recognition with sparse, localized features. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 11–18. IEEE, 2006. ISBN 0769525970.
- [28] A. Nordberg, J.O. Rinne, A. Kadir, and B. Långström. The use of PET in Alzheimer disease. *Nature Reviews Neurology*, 6(2):78–87, 2010. ISSN 1759-4758.
- [29] V. Rachakonda, PAN Tian Hong, and W.D. Le. Biomarkers of neurodegenerative disorders: How good are they? *Cell Research*, 14(5):349–358, 2004. ISSN 1001-0602.
- [30] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):77, 2011.
- [31] CM Roe, AM Fagan, MM Williams, N. Ghoshal, M. Aeschleman, EA Grant, DS Marcus, MA Mintun, DM Holtzman, and JC Morris. Improving CSF biomarker accuracy in predicting prevalent and incident Alzheimer disease. *Neurology*, 2011. ISSN 0028-3878.
- [32] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cor-

- tex. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 994–1000. IEEE, 2005. ISBN 0769523722.
- [33] M.C. Tartaglia, H.J. Rosen, and B.L. Miller. Neuroimaging in Dementia. *Neurotherapeutics*, pages 1–11. ISSN 1933-7213.
- [34] P.M. Thompson, K.M. Hayashi, G. De Zubicaray, A.L. Janke, S.E. Rose, J. Semple, D. Herman, M.S. Hong, S.S. Dittmer, D.M. Doddrell, et al. Dynamics of gray matter loss in Alzheimer’s disease. *Journal of Neuroscience*, 23(3):994, 2003.
- [35] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289, 2002. ISSN 1053-8119.
- [36] M. Van Herk. A fast algorithm for local minimum and maximum filters on rectangular and octagonal kernels. *Pattern Recognition Letters*, 13(7):517–521, 1992. ISSN 0167-8655.

An Evaluation of Sparse Inverse Covariance Models for Group Functional Connectivity in Molecular Imaging

4.1 Introduction

Neuroimaging modalities are routinely used to evaluate regional differences in brain function and connectivity. Neuroimaging provides researchers with a method to evaluate in-vivo brain function, where various statistics relating activity between spatially distributed regions are used as indicators of functional connectedness [10]. Many studies have shown changes in functional connectivity, as assessed with neuroimaging, across genders [4, 15] and in psychiatric illnesses such as schizophrenia [19, 22] and Alzheimer’s Disease [14, 36] among others. The majority of studies and methods of evaluating functional connectivity have focused on dynamic imaging modalities such as functional magnetic resonance imaging (fMRI) and $^{15}\text{O}\text{-H}_2\text{O}$ Positron Emission Tomography (PET). Fewer studies have evaluated functional connectivity in static molecular imaging modalities such as $^{18}\text{F}\text{-FDG}$ PET and Tc-99m HMPAO Single Photon Emission Computed Tomography (SPECT) across groups of subjects. Huang et al. [14] compares functional connectivity across cohorts using static

PET; however, their results give little information on how well the models recover the true connectivity profiles of the populations tested. Careful validation of connectivity models is necessary, prior to their use and interpretation in real data [29].

A relatively simple method of interrogating functional connectivity in neuroimaging is through inter-regional and/or voxel-wise correlations [6, 26]. This method typically consists of cross-correlating the mean functional activity for each region of interest (ROI) and applying a threshold to remove values close to zero. Unfortunately, it has been shown that the sample correlation matrix is often a poor estimator of the population correlation matrix and can result in incorrect conclusions based on estimation error [35] [29]. Further, correlations between two ROIs does not imply a direct functional connection and thus impedes our ability to interpret our results with respect to known brain circuits. A more interpretable statistic, that has been shown to estimate the true network [29], is the partial correlation: the statistical relationship between pairs of brain regions after removing the influence of all the others. If the brain regions are jointly Gaussian, then we can find regions with strong partial correlation by examining the non-zero entries in the scaled inverse covariance matrix of the Gaussian distribution, fit to the observed data. Unfortunately, in typical neuroimaging studies, we do not have enough data to accurately compute the inverse of the covariance matrix among brain regions; therefore, some form of regularization is needed. To estimate the inverse covariance matrix and thus the partial correlations consistent with our observed data, we can use regularized maximum likelihood methods [3, 31]. Although we have methods available, we lack data on how well these classes of models can recover the true connectivity in group-based molecular imaging and how sample sizes can effect accuracy of the results.

In this study we compare three models for learning functional connectivity: one where the connectivity profiles are learned independently for each cohort and two that share information between cohorts. In settings with low sample sizes, we hypothesize that models

which share connectivity information across cohorts will perform better in recovering correct connections. We develop a simple clustering model for defining functional regions of interest, used for functional connectivity, that combines anatomy from a stereotaxic atlas and functional activity from static molecular imaging scans. Our contribution is a quantitative assessment of how well the functional connectivity models can accurately recover a gold standard connectivity pattern across a range of typical data set sizes in neuroimaging studies based on a very large sample of SPECT scans. To our knowledge, this is the first study evaluating the accuracy of recovered connectivity profiles by sample size in group-based static molecular imaging. The rest of the manuscript is organized as follows. We first present the clustering model used to define the nodes of our network for use in functional connectivity. Next, we describe three models for learning sparse inverse covariance matrices which have been applied to similar problems. We conclude the methods section with a description of the gold standard connectivity data we use to quantitatively evaluate model performance. In the results section we present a detailed comparison of the sparse inverse covariance models with respect to a gold standard connectivity profile, as a function of sample size. We conclude with a discussion of the results and plans for future work.

4.2 Methods

In this section we describe three sparse inverse covariance models for learning group-based functional connectivity patterns between pairs of regions, one of which learns the patterns independently for each cohort and two models which share connectivity information across cohorts. We anticipate the models that share connectivity information across cohorts will improve our prediction accuracy in settings with small sample sizes. Prior to performing an analysis of functional connectivity, we must define the nodes of our network. In settings

where structural scans are unavailable, a popular choice is to use a standard atlas to define regions of interest (ROIs) such as the AAL atlas from Tzourio-Mazoyer et al. [34] and the functional signals in each anatomical boundary are averaged. Using mean functional signals from anatomically defined ROIs are prone to decreased sensitivity by averaging signals from a small number of activated voxels with noise [24]. In section 4.2.1 we take a pragmatic approach, given the data we have available for evaluating the inverse covariance models, and use finite mixture modeling to find functional clusters constrained by anatomically-defined boundaries for use as nodes in the network. We conclude the methods section with a description of the gold standard dataset we use to quantitatively evaluate the performance of each connectivity model.



4.2.1 Functional Clusters

Here we describe the finite mixture model for finding functional clusters within anatomically-defined regions of interest that will subsequently be used as nodes in our network for functional connectivity modeling. The model has been designed with respect to static SPECT and PET modalities, consisting of independent subjects across multiple cohorts, with no available anatomical information. Our motivation is to find a set of regions that are consistent with the observed functional data across all cohorts, while also allowing the nodes to be interpreted with respect to an anatomical reference. From neuroanatomy, we know the brain is fundamentally composed of spatially varying clusters of cells which segregate during development into specialized units. These specialized units consist of compact clusters of neurons, forming nuclei, which contribute to the overall function of the region. Neuroanatomic studies of brain cytoarchitectonics provides evidence for intra-regional specializations of neuronal clusters for many brain structures [12, 17, 27]. Using a simple average functional signal from an anatomically defined brain region over simplifies this complexity and risks missing information related to inter-regional variations in function.

To learn the location and number of functional nuclei in areas constrained by an anatomical atlas, we observe that if an anatomical region has many nuclei, we would expect to find multi-modal signals within the functional imaging data. Further, because functional imaging data is always smoothed with a Gaussian smoothing kernel during image formation, we would expect any signal from the nuclei, as measured by functional imaging, to be smoothly varying in space with a functional peak spatially close to the center-of-mass of the nuclei and an area roughly consistent with the extent of the cells composing the nuclei. Given these observations, we propose to model the functional clusters using finite Gaussian mixture models [21]. Based on data from neuroanatomic studies, there may be multiple nuclei within many anatomically-defined regions (e.g. Toncray and Krieg [33] discuss twenty-six nuclei of the human thalamus) but we do not know how many of these nuclei result in functional signals detectable in the acquired data. We can use this information to limit the number of clusters we search for and use model selection methods to identify how many clusters are needed to reasonably explain our observed functional signals.

We begin by thinking about the imaging data as being composed of a set of voxels, each being identified by its three-dimensional coordinate $x_i = (p_1, p_2, p_3)$ within the image. For a given voxel coordinate, x_i , the acquired functional imaging data, is quantified to counts per unit time and provides a non-negative, real-valued proxy $I(x_i) \in \mathbb{R}_{\geq 0}$ corresponding to the number of detected events by the imaging system at that voxel coordinate. We make the simplifying assumption that each detected event is independent from all other detections and identically distributed. This assumption implies that the functional imaging signal $I(x_i)$ at each voxel (i.e., the number of detected events) can be modeled as $I(x_i)$ independent observations at voxel x_i . For each region r defined in an anatomical atlas, the set of N^r voxel coordinates contained within the boundaries of the atlas region $r = \{x_1, \dots, x_{N^r}\}$ specify the voxels assigned to the anatomical region. Each voxel is a member of only one anatomical

region. The log-likelihood of all the detected events within a region is then modeled by a K-component Gaussian mixture model:

$$\ln p(X^r|\Theta^{(r)}) = \sum_{i=1}^{N^{(r)}} I(x_i) \ln \left\{ \sum_{k=1}^K w_k^{(r)} \mathcal{N}(x_i|\mu_k^{(r)}, \Sigma_k^{(r)}) \right\} \quad (4.1)$$

where $\Theta^{(r)} = \{w_1^{(r)}, \dots, w_k^{(r)}, \mu_1^{(r)}, \Sigma_1^{(r)}, \dots, \mu_k^{(r)}, \Sigma_k^{(r)}\}$ are the model parameters, X^r is the set of all voxel coordinates contained in region r , and $I(x_i)$ is the number of detected events at the voxel coordinate x_i for all $N^{(r)}$ coordinates. Parameters $\mu_k^{(r)}$ and $\Sigma_k^{(r)}$ represent the mean spatial location and covariance matrix of the k^{th} cluster in region r . The mixing weights $w_k^{(r)}$ represent the probability that x_i was generated by component k and are subject to the constraints $\sum_{k=1}^K w_k^{(r)} = 1$ and $0 \leq w_k^{(r)} \leq 1$ to be valid probabilities. The unknown parameters $\Theta^{(r)}$ can be learned using the iterative expectation maximization (EM) algorithm and maximized in closed form [8, 20]. Although the algorithm is guaranteed to converge, it is not guaranteed to converge to a global maximum of the likelihood function and is initialization dependent. To initialize the mixture model we use the kmeans++ algorithm of Arthur and Vassilvitskii [2], an extension to the K-means algorithm. K-means is a clustering technique which finds a partitioning of the data into K clusters that minimizes the sum of the squared distances between each data point and its closest cluster center and is routinely used to initialize Gaussian mixture models. The kmeans++ algorithm is a simple extension that uses a random seeding technique which has been shown to improve speed and accuracy with respect to the recovering the true cluster centers. To initialize our Gaussian mixture models using this technique, we choose an initial cluster center from the data with the highest number of detected events. We then choose the next cluster center at random from the remaining data points with probability proportional to the distance of each data point to the closest previously selected cluster center, weighted by the number of detected events. This step gives higher probability to data points farther away from the previously selected cluster centers, balanced with the frequency of detections at those points. The procedure

continues until all initial K cluster centers have been selected. We then run the k-means algorithm to convergence and use the assignments of the voxels to the K clusters to initialize our Gaussian mixture model means and covariances.

The final step in the clustering model is to select an appropriate setting for the number of clusters K for each region. We cannot simply choose the model that explains the data best because more complex models (i.e., larger K) will always explain the data better than a simpler model with fewer clusters. Among the various approaches to model selection, penalized likelihood methods have been shown to be competitive, simple to implement in practice, and to work well when applied to mixture models [30]. Here we choose the Bayesian information criterion (BIC) developed by Schwarz et al. [28] for model selection, which has been shown to be a consistent estimator for independent and identically distributed observations in linear exponential family models such as the Gaussian mixture model [11]. By choosing the model with the minimum BIC score, one is attempting to select the candidate model with the highest Bayesian posterior probability. The BIC score penalizes the model likelihood and tends to favor simple models over more complex ones. The BIC score for our mixture model within a brain region is given by:

$$BIC(\Theta^{(r)}) = -2 \ln p(X|\Theta^{(r)}) + \rho^{(r)} \ln \left(\sum_{i=1}^{N^{(r)}} I(x_i) \right) \quad (4.2)$$

where $\rho^{(r)}$ is the number of parameters in the model for region r and is used to penalize the likelihood relative to model complexity. In our model we are learning K three-dimensional Gaussian distributions for each region. For each Gaussian distribution of dimension $D = 3$, we need $D/2(D+1)$ parameters to specify the symmetric covariance matrix and D parameters to specify the mean. There are K mixing weights for the K -component Gaussian mixture model, resulting in $K * D/2(D + 3) + K$ parameters. After learning the mixture models for each setting of K , we compute the BIC score, penalizing the likelihood scores by the number

of parameters estimated. We select the best model as the one with the lowest BIC score, which implies the model with the highest relative data likelihood after penalizing for model complexity. Once the best models have been selected for each brain region, we compute values for each cluster by averaging over the number of detections at each voxel, weighted by their cluster membership probabilities. We use these cluster averages, computed for each subject, in the sparse inverse covariance models.

4.2.2 Sparse Inverse Covariance Estimation

In section 4.2.1 we described a clustering algorithm that parcellates the brain into functionally defined and anatomically constrained sub-regions. Ultimately we are interested in learning about how the brain is functionally connected at the group level across different experimental conditions and/or cohorts, in regions that both correspond to anatomy and have support from our functional data, because both aspects help us interpret the results. In learning about the functional connectivity in our data, we are presented with a few challenges. First, the brain is composed of approximately 100 billion neurons, with each neuron being connected with up to 10,000 other neurons, potentially forming 1 trillion synaptic connections. Even at the scale of neuroimaging, we are likely to find some association between any pair of brain regions. We need a principled approach that focuses our attention on the most relevant associations in a way that can be related to known brain circuits. Next, we are constrained by small sample sizes. Often, particularly in rare diseases, it is difficult to find large groups of subjects, prohibiting us from learning about functional connectivity. Fortunately, there are models available that can help us learn about functional connectivity in light of these problems, yet there is little data on how accurately the models perform when applied to molecular brain imaging data.

Sparse inverse covariance estimation models provide a structured approach of estimating

pairwise connections between variables in settings with small numbers of samples. Entries in the inverse covariance matrix (also called the precision matrix) of a Gaussian model, correspond to the pairwise statistical relationships between variables conditioned on all the other variables in the model. Values of zero in the inverse covariance matrix (i.e., zero partial correlation) imply conditional independence of the associated variables [18]. Alternatively, strong partial correlations are indicative of direct interactions, helping us to interpret results with respect to known brain circuits. In contrast, simple correlations provide pairwise associations but do not account for all the other brain regions and their potential effect on the correlation. Learning inverse covariance matrices in settings with small data set sizes and a large number of variables is difficult and some form of regularization is needed to both improve prediction accuracy and aid in interpretation, focusing on a smaller subset that exhibit the strongest effects [31]. In section 4.2.2.1 we describe Graphical Lasso (GL), a popular technique for learning sparse inverse covariance matrices. GL uses regularization to improve prediction accuracy and interpretability in settings with small sample sizes relative to the number of variables [31]; however, in many experiments the groups being compared are similar, and we would expect some of the functional connectivity relationships are shared across cohorts. We would like to use this observation to help improve our prediction accuracy. In section 4.2.2.2 we describe the Fused Graphical Lasso (FGL) and Group Graphical Lasso (GGL) models proposed by Danaher et al. [7] based on work from Tibshirani et al. [32] for jointly estimating sparse precision matrices across cohorts. Our contribution is a quantitative analysis of how well these models perform in recovering a gold standard connectivity profile as a function of sample size in static molecular imaging.

4.2.2.1 Graphical Lasso

The least absolute shrinkage and selection operator (lasso) proposed by Tibshirani [31] is a method for penalized regression that shrinks some parameter estimates and sets others to

zero. Shrinking parameter estimates controls over fitting and results in models that have better predictive accuracy. Setting some parameters to zero provides variable selection, focusing on the more relevant features of the data, resulting in more interpretable modeling results. Banerjee et al. [3] applied the lasso penalty to learning sparse undirected graphical models in a multivariate Gaussian setting and developed a block-wise interior point algorithm which they noted is equivalent to iteratively solving lasso problems. Friedman et al. [9] pursued this observation and developed the Graphical Lasso (GL) algorithm. The learning problem is to maximize the penalized log-likelihood over all positive definite matrices Φ given by:

$$\hat{\Sigma}^{-1} = \arg \max_{\Phi \succ 0} \log \det \Phi - \text{tr.}(S\Phi) - \lambda \|\Phi\|_1 \quad (4.3)$$

where $\hat{\Sigma}^{-1}$ is the estimate of the sparse precision matrix, S is the empirical covariance matrix of the data computed using the results from the clustering model in section 4.2.1, where entry $S_{i,j}$ is the covariance between clusters i and j across all subjects in the group, and $\|\Phi\|_1$ is the L_1 norm, the sum of the absolute values of the elements of the current estimate of Φ . The non-negative coefficient λ controls the relative importance of the sparsity-inducing L_1 regularization term. For each cohort, we compute the covariance matrix between all pairs of clusters and directly apply the GL algorithm to learn a precision matrix for each cohort independently. The GL algorithm cycles through the variables, fitting a modified lasso regression to each variable in turn. The algorithm is fast, enabling problems with thousands of parameters to be solved efficiently. Hsieh et al. [13] extended the graphical lasso algorithm, exploiting underlying structure in the data and optimizing computational bottlenecks, making it feasible to solve problems with billions of parameters on a single machine. In estimating functional connectivity across the entire brain, we can easily have thousands of parameters and small sample sizes, making this algorithm attractive.

4.2.2.2 Joint Graphical Lasso

In many neuroimaging experiments we are searching for subtle changes in functional relationships between a subset of brain areas, i.e., we are not expecting the relationships between all brain regions to be different across cohorts. In situations where we expect parameters to be shared across groups and have a small data set size relative to the number of parameters being estimated, it has been shown by Danaher et al. [7], using simulated data with a known amount of parameter sharing, that jointly estimated sparse precision matrices are closer to the true distributions than estimating them separately with the GL algorithm. The learning problem is similar to equation (4.3) except we replace the L_1 norm penalty with a generic penalty function $P(\{\Phi\})$ defined across the set of precision matrices $\{\Phi\}$:

$$\left[\left(\hat{\Sigma}^{(1)} \right)^{-1}, \dots, \left(\hat{\Sigma}^{(G)} \right)^{-1} \right] = \arg \max_{\Phi \succ 0} \left(\sum_{g=1}^G n_g [\log \det \Phi^{(g)} - \text{tr}(S^{(g)} \Phi^{(g)})] - P(\{\Phi\}) \right) \quad (4.4)$$

where $\left[\left(\hat{\Sigma}^{(1)} \right)^{-1}, \dots, \left(\hat{\Sigma}^{(G)} \right)^{-1} \right]$ are the estimates of the sparse precision matrices for all groups, n_g is the number of observations in group g , and the first term inside the sum is the contribution from each group to the log-likelihood. Danaher et al. [7] propose two convex penalty functions, the Fused Graphical Lasso (FGL) and Group Graphical Lasso (GGL) penalties. FGL encourages groups to share both network structure and parameter values; the FGL penalty is given by:

$$P(\{\Phi\}) = \lambda_1 \sum_{g=1}^G \sum_{i \neq j} \left| \Phi_{i,j}^{(g)} \right| + \lambda_2 \sum_{g < g'} \sum_{i,j} \left| \Phi_{i,j}^{(g)} - \Phi_{i,j}^{(g')} \right| \quad (4.5)$$

where λ_1 and λ_2 are non-negative coefficients controlling the amount of sparsity and similarity in the precision matrices across groups respectively. The first term in equation (4.5) is similar to the L_1 penalty in graphical lasso, except that the sum is over the off-diagonal elements. Increasing this penalty results in sparser networks. The second term is the absolute

difference in corresponding parameter estimates across the G precision matrices that we are estimating, one for each cohort. Increasing the λ_2 penalty will result in similar connectivity networks and strengths between pairs of variables across the groups. Making λ_2 extremely large will result in identical networks across the groups.

In contrast, the GGL penalty encourages a shared pattern of sparsity, without requiring similarity between the parameter estimates across the groups. The GGL objective is:

$$P(\{\Phi\}) = \lambda_1 \sum_{g=1}^G \sum_{i \neq j} |\Phi_{i,j}^{(g)}| + \lambda_2 \sum_{i \neq j} \left(\sum_{g=1}^G \Phi_{i,j}^{(g)2} \right)^{\frac{1}{2}} \quad (4.6)$$

where λ_1 and λ_2 are non-negative coefficients controlling the amount of sparsity and shared sparsity respectively. The first term in equation (4.5) is the same L_1 penalty as FGL; as before, making this penalty large results in sparser networks. The second term encourages a similar pattern of sparsity in the precision matrices across the groups. Interestingly, both penalty terms will contribute to sparsity. Danaher et al. [7] show GGL results in a weaker form of network similarity than FGL, where patterns of sparsity (e.g., functional connectedness) are similar across the groups, but the strength of connections in the network are not penalized for being different.

To estimate the networks using joint graphical lasso penalties, we use the alternating directions method of multipliers (ADMM) [5]. The algorithm is guaranteed to converge to the global optimum. For details on applying ADMM to this problem see Danaher et al. [7]. To apply the joint graphical lasso algorithms to our problem, we follow the same procedure as for graphical lasso except the precision matrices are learned jointly across cohorts. Compared to graphical lasso, the joint lasso penalties have two coefficients to set, λ_1 and λ_2 . In section 4.3 we will evaluate different ways of setting the penalties and whether the joint models improve accuracy in recovering the gold standard connectivity patterns and strengths across

a range of data set sizes.

4.2.2.3 Model Summary

The clustering and sparse inverse covariance algorithms described above are summarized graphically in Figure 4.1. The entire set of imaging data is collapsed across cohorts prior to clustering. For each region, `kmeans++` is run for each setting of K and used to initialize a Gaussian mixture model. The resulting clusters are evaluated using the BIC scores and the solution with the minimum BIC score for each region is selected and used as regions of interest. The region of interest averages are then calculated for each subject and each cluster across all regions. For each cohort, we then compute the empirical covariance matrix among all pairs of clusters and learn precision matrices using the graphical lasso, fused, and group graphical lasso algorithms. This process results in inverse covariance matrices with varying amounts of non-zero entries which are depicted in a graph structure by adding edge between pairs of variables. Figure 4.2 shows an example of the entire model run with eight anatomical regions (i.e., yellow region outlines) and voxels colored by their cluster membership within each anatomical region. The edges in the graph indicate non-zero entries in the inverse covariance matrix between pairs of clusters and are scaled by the magnitude of the partial correlations where blue edges indicate negative partial correlations and red areas positive partial correlations. Pairs of clusters that are not connected in the figure have zero entries in the inverse covariance matrix, reflecting their conditional independence. In the experiments discussed in section 4.3, these procedures were run in three-dimensions using the whole brain.

4.2.3 Gold Standard Data

Ultimately we are interested in learning about differences in how the brain is sharing information across groups of subjects, evaluated with static molecular imaging techniques, because

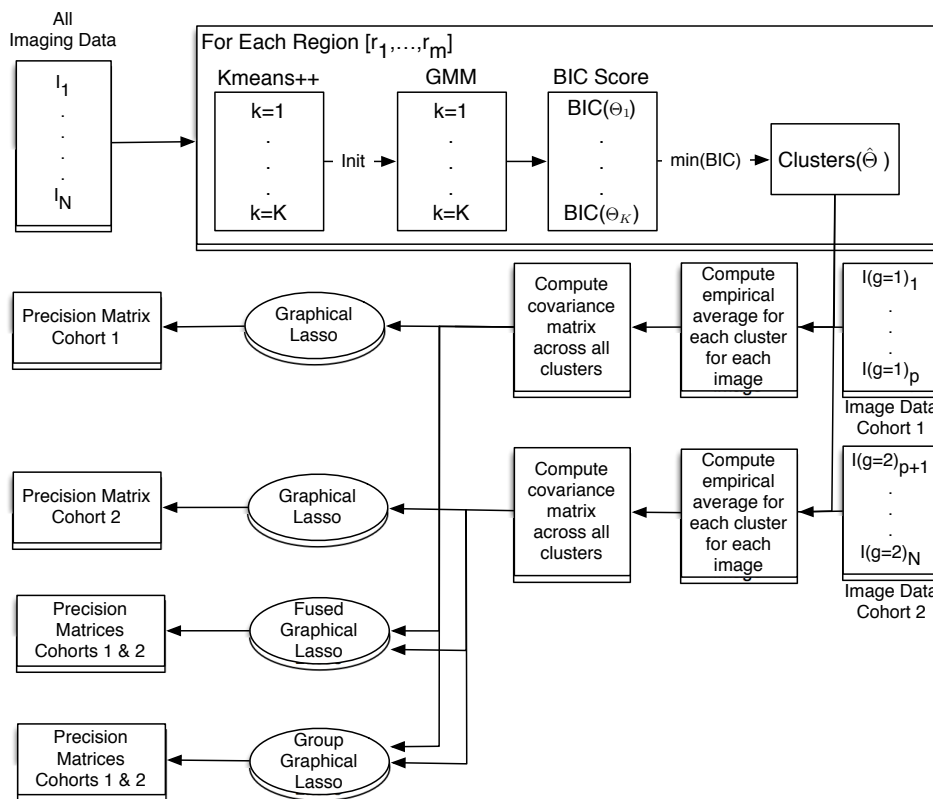


Figure 4.1: Summary of the methods for learning functional clusters and precision matrices using graphical lasso, fused graphical lasso, and group graphical lasso algorithms in group-based static molecular imaging data. Note, in this example there are only two cohorts but the process is the same for additional groups.

it may provide important information on overall brain function. In practical settings, the amount of available data is small relative to the number of parameters needed to estimate whole-brain connectivity, which could result in both spurious connections and/or poor estimates of true connections due to dependencies in the parameters [35]. In section 4.2.2 we described three models for functional connectivity that use regularization to, hopefully, improve predictive accuracy and reduce variance in our parameter estimates, at the expense of some bias, while also making the results more interpretable, focusing our attention on the strongest functional connections [31]. In order to evaluate the performance of the inverse covariance models in finding true functional connections in molecular imaging data, it is necessary to have a large data set with many samples relative to the number of parameters

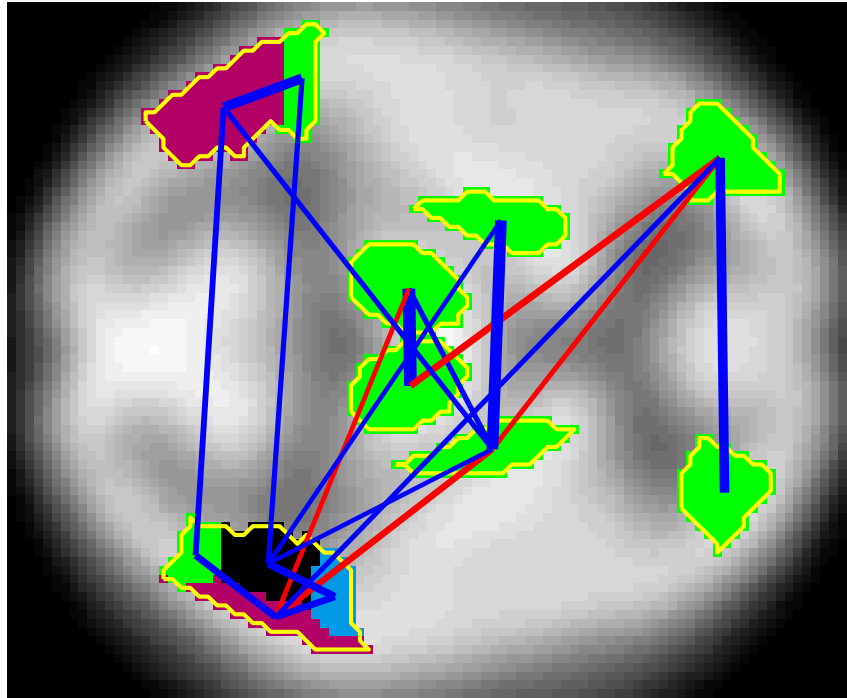


Figure 4.2: Example of clustering and inverse covariance modeling in a two-dimensional SPECT scan slice. Yellow outlines correspond to anatomical regions from an atlas, voxels are colored by their cluster membership within each anatomical region, and lines connecting clusters correspond to the non-zero parameters in the inverse covariance matrix between pairs of clusters and are scaled by the magnitude of the partial correlations where blue edges indicate negative partial correlations and red areas positive partial correlations.

we are estimating.

The Amen Clinics Inc. (<http://www.amenclinics.com/>) has been collecting technetium-99m hexamethylpropyleneamine oxide (Tc99m HMPAO) SPECT scans for many years on a variety of disorders such as attention deficit hyperactivity disorder, depression, anxiety, and behavioral problems, among others. The Amen Clinics Inc. provided us with a SPECT data set to use as a gold standard, consisting of 11,906 males (avg. age 29.2 ± 17.6) and 7,550 females (avg. age 35.6 ± 18.4) [1]. This is the largest SPECT data set that we know

of. On average, 55% of the subjects have mood disorders, 7% bipolar, 25% depression, 47% ADHD, 45% anxiety, 10% substance abuse, and 32% brain trauma. Although the data set consists of a variety of disorders and comorbidities, for this purpose we need two or more large cohorts that have both shared and unshared functional connections, to evaluate how well the three inverse covariance models recover these patterns as a function of sample size.

Subjects in the Amen data set were injected with an age/weight appropriate dose of Tc99m HMPAO and were at rest during uptake. All subjects were scanned on a high-resolution Picker Prism 3000 triple-headed gamma camera with fan beam collimators. The original reconstructed image matrices were 128x128x29 voxels with sizes of 2.16mm x 2.16mm x 6.48mm and values representing counts. The images were spatially normalized to the MNI atlas using SPM8 software [23], resulting in image matrices of 79 x 95 x 68 voxels in x, y, and z dimensions respectively with isotropic 2mm voxel sizes. The Automated Anatomical Labeling (AAL) atlas was used to define the brain regions based on the anatomical parcelations available in the atlas because no structural imaging data was available [34].

After normalization, the clustering model (section 4.2.1) for $K \in \{1, \dots, 25\}$ was run across the entire data set, for each region in the AAL atlas, except the cerebellum. The cerebellum was not included due to missing data in the lower slices of some scans. The clustering model, after selecting the best model for each region using the BIC scores, resulted in 180 clusters across the brain as shown in Figure 4.3. The cluster averages for each image, stratified by cohorts, were computed and mean centered. The empirical correlation matrices for both cohorts are shown in Figure 4.4. Although the sample is large relative to the number of variables, there is still measurement noise in the data. To remove some of this noise, we run the graphical lasso algorithm with light regularization ($\lambda = 0.1$), resulting in precision matrices with 2119 pairwise connected regions in males, 2130 pairwise connections in females, and 1750 shared connections ($\approx 82\%$) across both cohorts. These precision matrices will

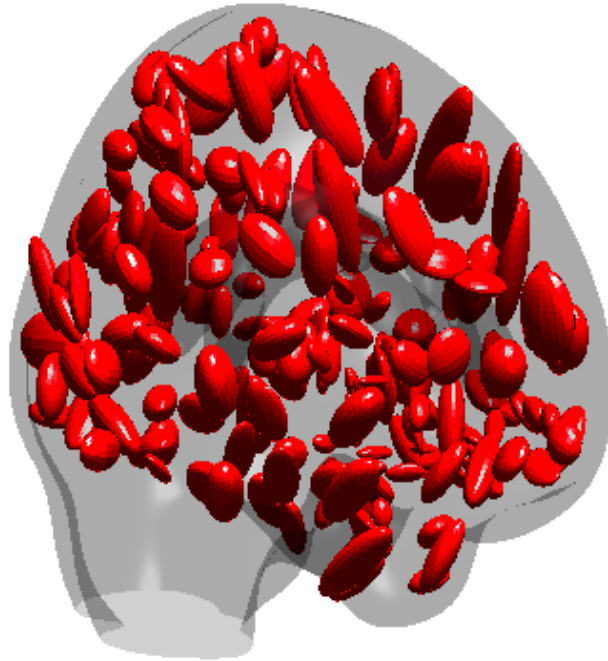
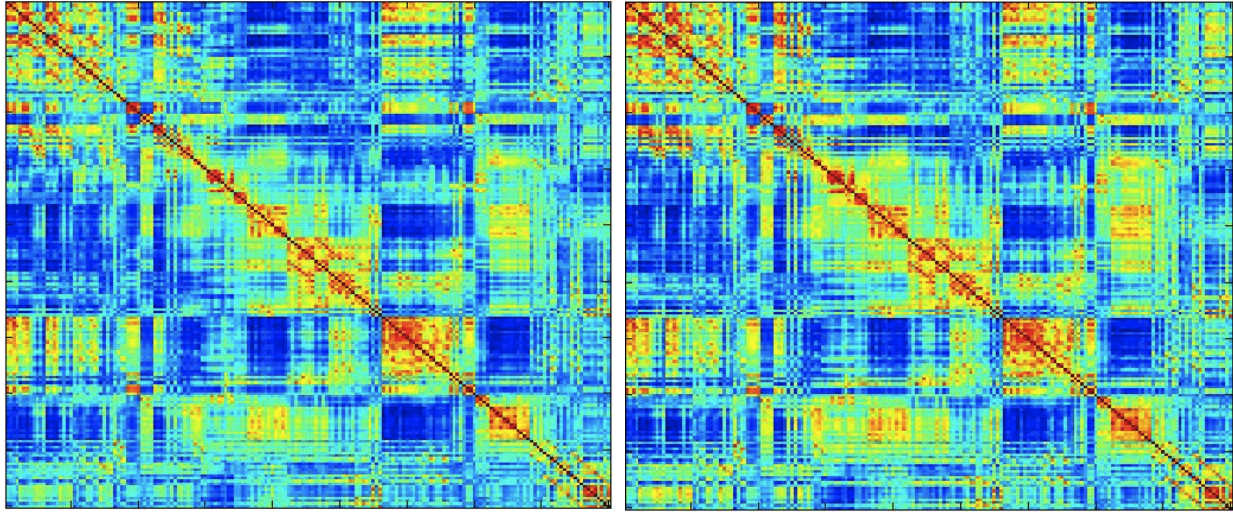


Figure 4.3: Results from the clustering model run on the gold standard SPECT data set ($n=19,456$) collapsed across groups. Ellipsoids show the ± 1 standard deviation region around each cluster mean.

be used as a gold standard in Section 4.3 to quantitatively evaluate the performance of the inverse covariance estimation models in recovering the patterns as a function of the amount of available training data.

4.3 Results

In this section we provide a quantitative evaluation of the sparse inverse covariance models in terms of correct connections and connection strengths, across different sample sizes compared to the gold standard precision matrices described in section 4.2.3. The intent of our experiments are to evaluate the performance of the inverse covariance models at sample sizes typically used in functional imaging studies. Each of the models described in section 4.2.2 require a setting for the amount of regularization, which affects the sparsity of the results and the overall network structure; therefore, we evaluate two methods of setting the regu-



(a) Males

(b) Females

Figure 4.4: Empirical correlation matrices computed from the average values for each cluster for (a) males ($n=11,906$) and (b) females ($n=7,550$). Correlations range from -0.5 to 1.0

larization weights. In section 4.3.0.1 we use the gold standard precision matrices directly to determine optimal settings for the λ_2 regularization coefficients in the FGL and GGL models and compare receiver operating characteristic (ROC) curves of true-positive (TP) and false-positive (FP) connections across all models by varying λ in GL and λ_1 in FGL and GGL. In practice, one does not typically have access to a gold standard. In section 4.3.0.2 we use cross-validation to determine the regularization weights and compare the results with those determined using the gold standard.

4.3.0.1 Using Gold Standard for Parameter Settings

In this experiment we evaluate the sparse inverse covariance models by sample size, using the gold standard precision matrices to determine settings for the regularization parameters. To compare GL, FGL, and GGL models, we follow a procedure similar to Danaher et al. [7]. We fix the regularization coefficients of the group penalties (i.e., λ_2) and varying the λ_1 penalty in the FGL and GGL models and the λ penalty in the GL model, creating ROC curves of model accuracy in predicting correct functional connections. Further, we compare

the strength of connections relative to the gold standard precision matrices. To determine fixed settings for the group penalties, we use the following procedure. We first draw two random samples ($n=500$) from the gold standard data set, stratified by group. We then compute the cluster averages using the clusters found for the gold standard data set described in section 4.2.3. Using these data, we perform a grid search over the λ_1 and λ_2 parameter space, for FGL and GGL models separately, comparing the learned precision matrices ($\hat{\Phi}$) to those of the gold standard (Φ^*) in terms of their L_1 difference ($\sum_{g=1}^G |\hat{\Phi}_g - \Phi_g^*|$). We select the settings that result in the minimum L_1 difference from the gold standard. For the FGL model, the minimum was found with settings $\lambda_1 = 0.01$ and $\lambda_2 = 0.1$. For the GGL model, the minimum was found with settings $\lambda_1 = 0.001$ and $\lambda_2 = 0.01$. We will use these fixed settings for λ_2 in our comparison of the group regularized models (i.e., FGL and GGL) with GL.

Next, we draw 10 subsets of images, with replacement, stratified by groups, with equal sizes

$N \in \{10, 25, 50, 100, 250, 500, 750, 1000, 1500, 2000, 2500\}$ from the gold standard data set, after removing the subjects used in the parameter searching routine described above. Using the clusters found for the gold standard data set described in section 4.2.3, we compute the cluster averages and mean center the data. Next, we learn precision matrices using the GL model across a wide range of λ regularization penalties. Similarly, using the fixed λ_2 settings, we learn precision matrices using the FGL and GGL models across the same range of λ_1 regularization settings as GL.

The results showing the average number (across both groups and 10 subsets) of true-positive edges (TP) correctly identified (i.e., non-zero entries in the test precision matrices match those from the gold standard), compared to the average number of false-positive edges (FP) incorrectly identified (i.e., non-zero entries in the test precision matrices unmatched with the gold standard) by sample sizes are shown in Figures 4.5-4.7. The colored regions around

each marker indicate the unit standard deviation region about the means. The large markers identify points on the curves with the lowest sum of squared errors (SSE) between the test precision matrices and the gold standard for each model, indicating the point on the curves where connection strengths in the test precision matrices best match the gold standard. Interestingly, these occur in regions with a high proportion of true-positive edges relative to false positive edges. Visually, the FGL and GGL model curves generally dominate the GL curves for most sample sizes. The biggest difference between the joint models and graphical lasso are in the moderate sample sizes $N \in \{25, 50, 100\}$, where we see the largest benefit from parameter sharing across the groups. Both joint models, in general, perform equally well in terms of the presence or absence of edges. The average areas under the TP/FP edge curves (AUC) across the subsets, by sample sizes, are shown in Table 4.1 (columns 2-4). The T-test and Bonferroni corrected p-values, comparing all combinations of the models by sample sizes, are shown in columns 5-7. The AUCs are statistically lower in the GL model (i.e., negative t-scores) than both the FGL and GGL models at data set sizes less than 1000, with the moderate sample sizes showing the largest decreases. At data set sizes above 2000, the GL model outperforms the joint models, which is related to having enough data to accurately estimate the precision matrices without the added benefit of parameter sharing in the joint models, and the fact that we used the GL algorithm to remove noise when creating the gold standard precision matrices.

To evaluate how well the entries in the precision matrices corresponding to strength of functional connections match those of the gold standard data set, we calculate the SSE between the test precision matrices and the gold standard by model and sample size. The average minimum SSEs (across both groups and 10 subsets) are shown in Table 4.2 (columns 2-4). The T-test and Bonferroni corrected p-values comparing all combinations of the models by sample sizes are shown in Table 4.2 (columns 5-7). The results are similar to the TP/FP edge results, where the joint models perform better than the GL model at moderate sample

sizes, with the largest difference at $N = 100$. Between the FGL and GGL joint models, the FGL model performs better, yielding lower SSEs compared to the gold standard. Unlike the TP/FP edge results, the FGL model achieves a lower SSE than the GL model, even for the larger data sets (i.e $N \in 1500, 2000, 2500$), whereas GGL does not. To evaluate the effect of the clustering model on these results, we repeated our experiments by averaging all the voxels within each region of interest instead of using the clustering model. The results from the three sparse inverse covariance models are shown in Appendix Tables A.1 and A.2 and are consistent with those presented in this section.

Table 4.1: TP/FP Edge Area Under Curve (AUC) T-test Comparison by Model and Sample Size using Gold Standard for Parameter Settings

	GL	FGL	GGL	GL-FGL	GL-GGL	FGL-GGL
N	$AUC \pm std$	$AUC \pm std$	$AUC \pm std$	T(P Bonferroni)	T(P Bonferroni)	T(P Bonferroni)
2500	0.820±0.003	0.790±0.008	0.810±0.008	10.74(1.15E-07)	3.93(3.79E-02)	-5.51(1.22E-03)
2000	0.812±0.004	0.797±0.015	0.806±0.013	2.88(3.87E-01)	1.31(2.08E-01)	-1.33(1.00e00)
1500	0.799±0.004	0.792±0.018	0.806±0.013	1.19(1.00e00)	-1.44(1.00e00)	-1.87(1.00e00)
1000	0.784±0.004	0.799±0.005	0.811±0.003	-8.14(7.52E-06)	-18.18(1.93E-11)	-6.70(1.09E-04)
750	0.769±0.004	0.789±0.007	0.797±0.005	-8.26(6.07E-06)	-14.16(1.31E-09)	-2.79(4.68E-01)
500	0.683±0.006	0.776±0.006	0.779±0.004	-36.99(7.65E-17)	-45.47(1.93E-18)	-1.27(1.00e00)
250	0.660±0.003	0.746±0.007	0.745±0.007	-35.21(1.83E-16)	-35.30(1.75E-16)	0.13(1.00e00)
100	0.631±0.006	0.696±0.006	0.695±0.006	-24.07(1.50E-13)	-23.26(2.73E-13)	0.21(1.00e00)
50	0.604±0.004	0.652±0.005	0.653±0.005	-22.95(3.44E-13)	-24.43(1.16E-13)	-0.45(1.00e00)
25	0.572±0.004	0.621±0.013	0.626±0.012	-11.28(5.28E-08)	-13.19(4.26E-09)	-0.85(1.00e00)
10	0.537±0.007	0.601±0.013	0.604±0.013	-13.24(3.98E-09)	-14.30(1.12E-09)	-0.58(1.00e00)

Table 4.2: Sum of Squared Errors (SSE) with Gold Standard by Model and Sample Size

	GL	FGL	GGL	GL-FGL	GL-GGL	FGL-GGL
N	$SSE \pm std$	$SSE \pm std$	$SSE \pm std$	T(P Bonferroni)	T(P Bonferroni)	T(P Bonferroni)
2500	1.35±0.27	1.34±0.18	1.74±0.16	0.11(1.00E+00)	-3.93(3.90E-02)	-5.26(3.90E-03)
2000	1.66±0.23	1.57±0.20	2.08±0.30	0.96(1.00E+00)	-3.56(8.58E-02)	-4.54(1.17E-02)
1500	2.16±0.25	1.90±0.21	2.33±0.22	2.52(8.35E-01)	-1.66(1.00E+00)	-4.55(7.80E-03)
1000	3.29±0.25	2.70±0.41	3.09±0.30	3.85(4.68E-02)	1.59(1.00E+00)	-2.44(9.95E-01)
750	4.32±0.37	3.47±0.43	3.92±0.53	4.77(7.80E-03)	1.94(1.00E+00)	-2.11(1.00E+00)
500	6.03±0.68	4.78±0.67	5.15±0.44	4.14(2.34E-02)	3.42(1.17E-01)	-1.49(1.00E+00)
250	11.41±1.34	8.92±1.14	8.68±0.76	4.46(1.17E-02)	5.61(3.90E-03)	0.56(1.00E+00)
100	67.51±18.79	22.58±1.69	21.20±1.40	7.53(3.90E-03)	7.77(3.90E-03)	1.98(1.00E+00)
50	64.34±7.46	60.42±6.83	57.70±4.71	1.23(1.00E+00)	2.38(1.00E+00)	1.04(1.00E+00)
25	62.98±5.59	58.46±5.74	62.70±5.84	1.78(1.00E+00)	0.11(1.00E+00)	-1.64(1.00E+00)
10	100.52±10.29	83.90±4.48	84.13±4.49	4.68(7.80E-03)	4.62(7.80E-03)	-0.11(1.00E+00)

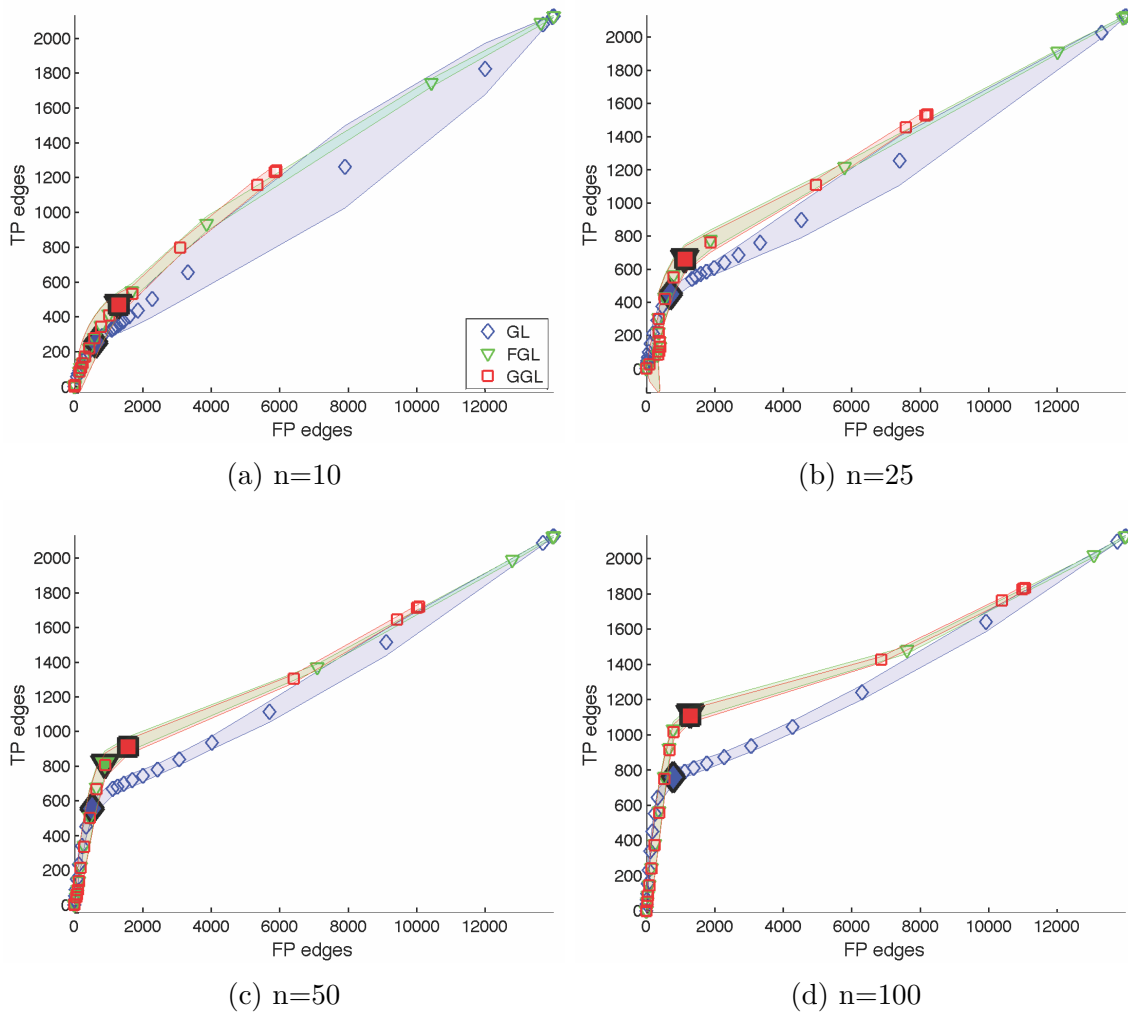


Figure 4.5: True-positive (TP) vs. false-positive (FP) edges between the test precision matrices and the gold standard for sample sizes $n \in \{10, 25, 50, 100\}$ using GL, FGL, and GGL inverse covariance models. Small markers indicate the means across the 10 test subsets and both cohorts. Colored regions show the unit standard deviation about the means. Large markers identify points with lowest average SSE compared to gold standard.

🧠 4.3.0.2 Using Cross-Validation for Parameter Settings

In section 4.3.0.1 we used the precision matrices from the gold standard data set directly in determining fixed group regularization penalties (i.e., λ_2) for the joint models. In practice, one does not typically have access to a gold standard. In this experiment we consider how well these models perform in a practical setting, determining the regularization coefficient settings using a typical 10-fold cross validation design. For this experiment we require 10 subsets of

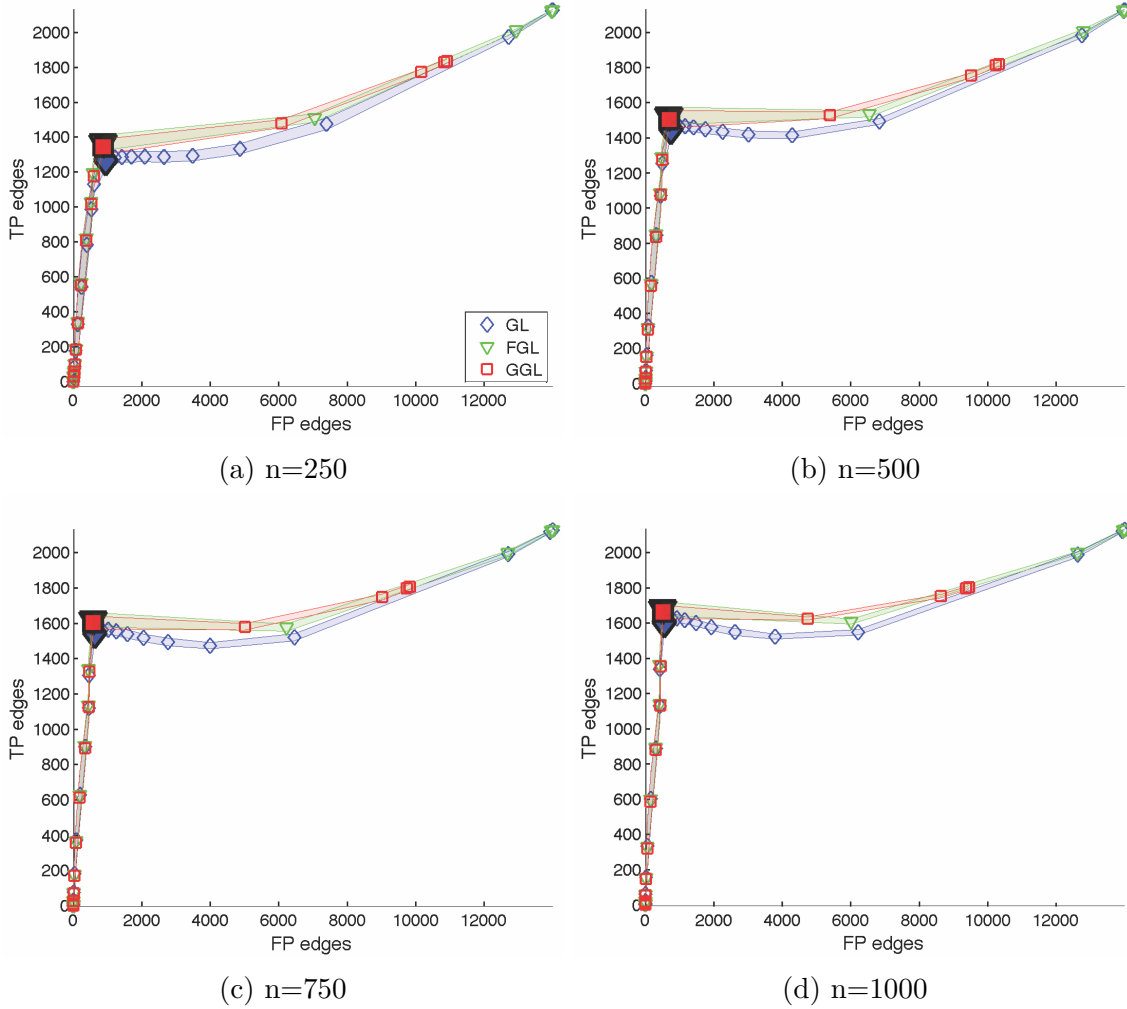


Figure 4.6: True-positive (TP) vs. false-positive (FP) edges between the test precision matrices and the gold standard for sample sizes $n \in \{250, 500, 750, 1000\}$ using GL, FGL, and GGL inverse covariance models. Small markers indicate the means across the 10 test subsets and both cohorts. Colored regions show the unit standard deviation about the means. Large markers identify points with lowest average SSE compared to gold standard.

data, stratified by group, for clustering, regularization coefficient selection, and testing with the gold standard. We first separate the gold standard data set into three independent sets, stratified by group ($N^{(g^1)} = 3,802$, $N^{(g^2)} = 2,516$ in each set). From each set and group, we draw 10 subsets of images, with replacement, of equal sizes $N \in \{10, 25, 50, 100, 250, 500\}$. This procedure results in 10 subsets of data for clustering, orthogonal to the 10 subsets for regularization coefficient selection, orthogonal to the 10 subsets for precision matrix evalua-

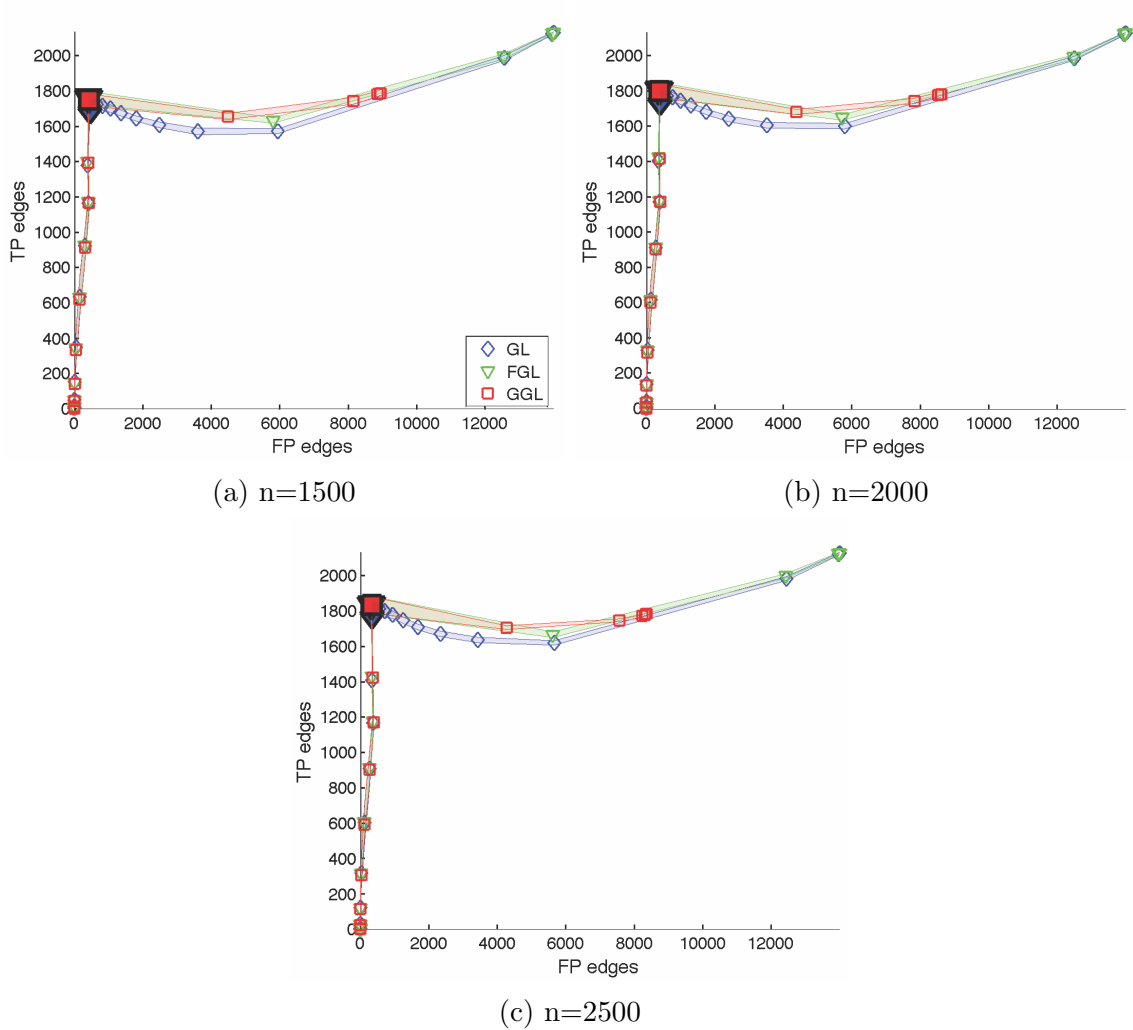


Figure 4.7: True-positive (TP) vs. false-positive (FP) edges between the test precision matrices and the gold standard for sample sizes $n \in \{1500, 2000, 2500\}$ using GL, FGL, and GGL inverse covariance models. Small markers indicate the means across the 10 test subsets and both cohorts. Colored regions show the unit standard deviation about the means. Large markers identify points with lowest average SSE compared to gold standard.

tion. Next, we run the clustering model described in section 4.2.1 on each of the clustering subsets independently, collapsing across cohort as done previously. Using the clusters learned from each clustering subset, we compute the average values for each cluster and each subject in the corresponding regularization coefficient selection subset and compute the corresponding covariance matrices between each of the clusters. We now have 20 covariance matrices, one for each of the 10 regularization coefficient selection subsets, for each cohort, using the

clusters learned from orthogonal data. These covariance matrices will be used to determine settings for the regularization coefficients. To set the regularization coefficients, we perform grid searches over the regularization parameter spaces (i.e., λ_1 , λ_2 for FGL and GGL, and λ for GL) for each inverse covariance model, selecting the regularization coefficient settings that minimizes the BIC score from each regularization coefficient selection subset, where the number of parameters are the number of non-zero edges in the precision matrices. After performing this procedure, we have settings for the regularization coefficients for each of the three inverse covariance models, for each of the 10 subsets of data. Finally, we use the clusters and the regularization coefficients to compute the precision matrices for each of the 10 precision matrix evaluation subsets and both cohorts, for each of the three inverse covariance models.

The results from this experiment are shown in Figure 4.8 and Table 4.3. Figure 4.8 shows the average number of TP and FP connections, by model, across the 10-folds compared to the gold standard data set. The lower segment of the stacked bars (i.e., lighter color) show the number of FP connections whereas the upper segment (i.e., darker color) show the number of TP connections. For $N \in \{100, 250, 500\}$ the FGL and GGL models find more TP connections and less FP connections than GL. For $N \in \{10, 25, 50\}$ GL finds more TP connections but also has more FP connections, except in $N = 50$ where both FGL and GGL models have more FP connections than GL. The SSE between the average precision matrices across the 10 test subsets and the gold standard are shown in Table 4.3. The results are similar to those of the TP/FP edges for $N \in \{100, 250, 500\}$, where FGL and GGL models achieve statistically lower SSEs than the GL model. In contrast, at the lower N 's, there is little significant difference between the models except in $N = 50$ where the GL model statistically outperforms the GGL model in correct connection strengths. Interestingly, the average SSE measurements for the GL model in the larger groups (i.e., $N \in \{100, 250, 500\}$) are much higher than in the experiment described in section 4.3.0.1; whereas, in the FGL

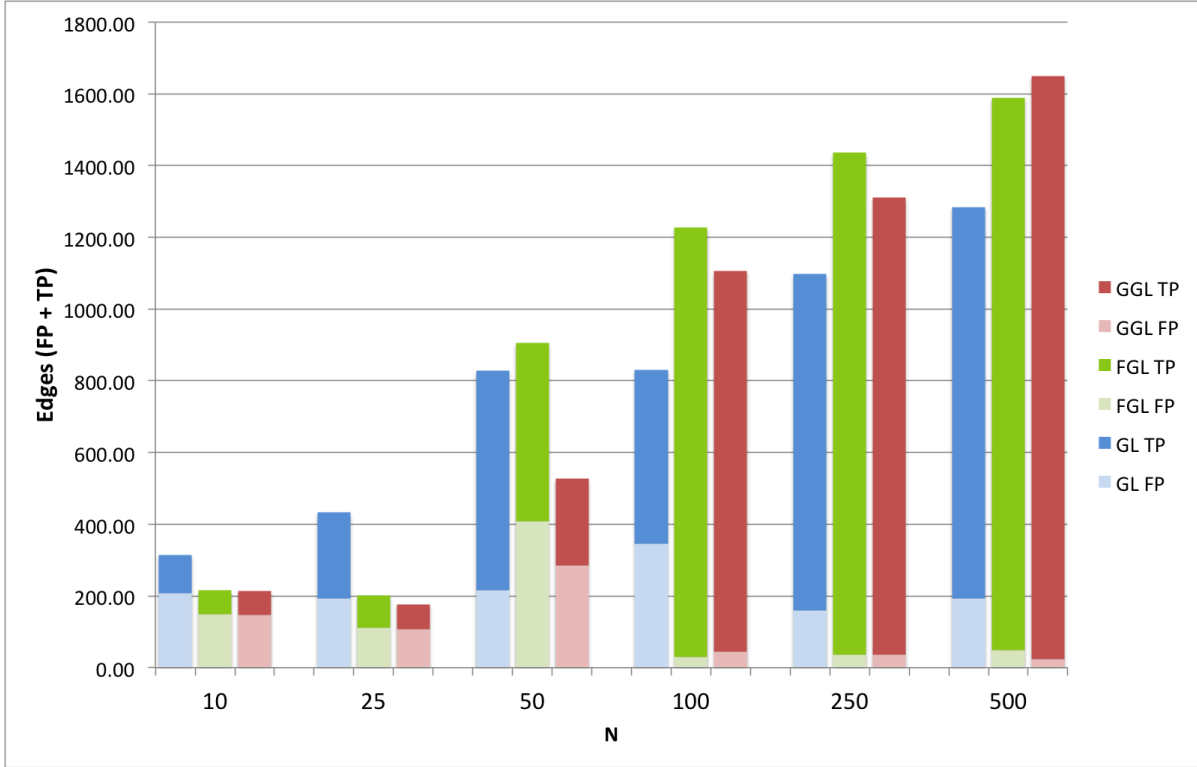


Figure 4.8: TP/FP edges by sample size and model for cross-validation experiment. Lower, light colored segments show counts of FP edges whereas, upper, dark colored segments show counts of TP edges.

and GGL models the results are fairly consistent. In looking at the regularization parameter settings across the splits of the data for this experiment, we find the GL model is selecting higher regularization parameters (range 0.2 – 0.7 depending on split and number of samples) as compared to the settings yielding the minimum average SSE when using the gold standard to set them directly (range 0.1 – 0.3). Conversely, in the FGL and GGL models, we find the λ_1 regularization parameter settings to be close to those found yielding the minimum average SSE when using the gold standard; whereas, the λ_2 , group regularization setting is generally a bit higher. Because the group models use all the data, across both cohorts, in setting the regularization parameters as compared to the GL model, the settings are more stable across splits and result in better estimates of the connection strengths as demonstrated in Table 4.3.

Table 4.3: Sum of Squared Errors (SSE) with Gold Standard by Model and Sample Size Using Cross-Validation

	GL	FGL	GGL	GL-FGL	GL-GGL	FGL-GGL
N	$SSE \pm std$	$SSE \pm std$	$SSE \pm std$	T(P Bonferroni)	T(P Bonferroni)	T(P Bonferroni)
500	156.66±70.11	4.88±1.04	8.54±3.32	6.85(3.77E-05)	6.67(5.26E-05)	-3.33(6.76E-02)
250	190.47±41.17	7.16±0.96	11.46±1.56	14.08 (6.69E-10)	13.74(9.98E-10)	-7.42(1.26E-05)
100	266.49 ±75.68	14.28±1.64	31.82±16.17	10.54(7.15E-08)	9.59(3.05E-07)	-3.41(5.59E-02)
50	207.17±46.93	231.15±128.5	332.05±91.61	-0.55(1.00E+00)	-3.84(2.18E-02)	-2.02(1.00E+00)
25	283.22±68.10	366.5±55.47	372.8±49.69	-3.00(1.39E-01)	-3.36(6.27E-02)	-0.27(1.00E+00)
10	530.03±167.41	342.01±41.8	350.91±45.33	3.45(5.19E-02)	3.27(7.73E-02)	-0.46(1.00E+00)

4.4 Discussion

In section 4.3 we described two experiments to evaluate the inverse covariance models described in section 4.2. In section 4.3.0.1 we used the gold standard precision matrices to set the group regularization parameters in the joint models that yielded the smallest difference from the gold standard. The results suggest that the joint models perform better in detecting true-positive functional connections and more accurately learn the connection strengths across all of the sample sizes. At $N \in \{1500, 2000, 2500\}$ the joint models find $\sim 2 - 3\%$ more true positive edges than the GL algorithm, at $N \in \{250, 500, 750, 1000\}$ the joint models find $\sim 3 - 7\%$ more true positive edges than GL, and at $N \leq 100$ the joint models find $\sim 31 - 46\%$ more true positive edges than the GL algorithm. The largest gain for the joint models (i.e., FGL and GGL) over the independent model (i.e., GL) in our comparisons were at $N = 100$. In section 4.3.0.2 we used a cross-validation design and orthogonal splits of the gold standard data to evaluate the performance of these models in a more practical setting, where a gold standard is not available. The results of this experiment similarly show the joint models performing better than the GL model for data set sizes of 100 or more, with the largest difference at $N = 100$. At $N \in \{500, 250, 100\}$ the joint models find $\sim 26 - 59\%$ more true positive edges than the GL algorithm. At $N = 50$ the FGL and GL algorithms detect only $\sim 23\%$ and $\sim 28\%$ of the true positive edges respectively; whereas, the GGL algorithm is much lower at $\sim 11\%$ of the true positive edges. Even worse, at $N \leq 25$

the true positive rates for all models are $\sim 3 - 4\%$ which is far too low for practical use. Further, we observe that the GL model SSE measures are much higher across the range of data set sizes when using cross validation to set the regularization weight than when using the gold standard directly (section 4.3.0.1), caused by over regularization of the GL model in the cross-validation experiment. The results at sample sizes less than 50 are generally quite poor and none of the models perform well using cross-validation and BIC scores to select settings for the regularization parameters. Visually, we see the GL model with more true-positive edges at $N \in \{10, 25\}$ than the joint models; although, all models appear to be over regularizing. From the results of experiment 1, using the gold standard to set the regularization parameters, we know the joint models can perform well at the smallest sample sizes, recovering $\sim 21 - 22\%$ of the true positive edges at $N = 10$ versus $\sim 11\%$ for the GL model, yet it is evident that care should be taken when setting the regularization parameters in the cross-validated design with such small samples. It may be prudent to compute the inverse covariance models across a range of regularization parameters and interpret the results based on edge selection frequencies or by controlling for false discovery rates as in Liu et al. [16] or to use a larger hold-out sample to set the regularization parameters prior to applying them to very small datasets.

Our results generally agree with Danaher et al. [7] who showed an improvement in the accuracy of the joint models over graphical lasso in groups with shared patterns of edges in simulated data. In our experiments, using molecular imaging data, both joint models perform about equally well. There does not seem to be an obvious advantage in choosing between FGL and GGL models in our experiments. In practice, one should take into consideration the expected similarity of the groups being compared, and the amount of available data, to determine whether the added complexity of selecting two regularization coefficients in the FGL and GGL models is warranted.

4.5 Conclusions

In this chapter we have compared models for group-based functional connectivity using static molecular imaging data and given a quantitative evaluation of the models in recovering a gold standard connectivity profile, as a function of sample size. In the largest samples, all models performed well; although, the joint models generally perform better. In smaller samples, the joint models are more stable and achieve better results when using a large dataset to determine settings for the regularization coefficients. Caution should be used when applying these models to very small data sets. Our experiments suggest the true positive rates will be low and the connection strengths will be inaccurate when using cross-validation to set the regularization parameters. Given these caveats, our results show there is value in using sparse inverse covariance estimation for measuring functional connectivity in group-based molecular imaging. It would be interesting to extend these experiments to within-subject experimental designs. In recent work, Qiu et al. [25] has developed a joint penalty that appears to perform better than GGL in settings where there are dependencies between networks. Using experiments similar to those presented in this manuscript, an evaluation by sample size in within-subject designs would provide complimentary information on using sparse inverse covariance estimation for functional connectivity modeling in molecular imaging.

Bibliography

- [1] Daniel Amen, Manuel Trujillo, David Keator, Derek Taylor, and Kristen Willeumier. Gender differences in regional cerebral blood flow in a healthy and psychiatric cohort of 46,034 spect scans. *The Open Neuroimaging Journal*, Under Review.
- [2] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.
- [3] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d’Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.
- [4] Bharat B Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M Smith, Christian F Beckmann, Jonathan S Adelstein, Randy L Buckner, Stan Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010.
- [5] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.
- [6] CM Clark, R Kessler, MS Buchsbaum, RA Margolin, and HH Holcomb. Correlational methods for determining regional coupling of cerebral glucose metabolism: A pilot study. *Biological Psychiatry*, 1984.
- [7] Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse

- covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.
- [8] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.
- [9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- [10] KJ Friston. Functional and effective connectivity in neuroimaging: a synthesis. *Human brain mapping*, 2(1-2):56–78, 1994.
- [11] Dominique MA Haughton et al. On the choice of a model to fit data from an exponential family. *The Annals of Statistics*, 16(1):342–355, 1988.
- [12] T Hirai and EG Jones. A new parcellation of the human thalamus on the basis of histochemical staining. *Brain Research Reviews*, 14(1):1–34, 1989.
- [13] Cho-Jui Hsieh, Mátyás A Sustik, Inderjit Dhillon, Pradeep Ravikumar, and Russell Poldrack. Big & quic: Sparse inverse covariance estimation for a million variables. In *Advances in Neural Information Processing Systems*, pages 3165–3173, 2013.
- [14] Shuai Huang, Jing Li, Liang Sun, Jun Liu, Teresa Wu, Kewei Chen, Adam Fleisher, Eric Reiman, and Jieping Ye. Learning brain connectivity of alzheimer’s disease from neuroimaging data. In *NIPS*, volume 22, pages 808–816, 2009.
- [15] Madhura Ingalhalikar, Alex Smith, Drew Parker, Theodore D Satterthwaite, Mark A Elliott, Kosha Ruparel, Hakon Hakonarson, Raquel E Gur, Ruben C Gur, and Ragini Verma. Sex differences in the structural connectome of the human brain. *Proceedings of the National Academy of Sciences*, 111(2):823–828, 2014.

- [16] Weidong Liu et al. Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978, 2013.
- [17] Jürgen K Mai, Joseph Assheuer, and George Paxinos. *Atlas of the human brain*. Academic Press San Diego:, 1997.
- [18] Guillaume Marrelec, Alexandre Krainik, Hugues Duffau, Mélanie Pélégrini-Issac, Stéphane Lehericy, Julien Doyon, and Habib Benali. Partial correlation for functional brain interactivity investigation in functional mri. *Neuroimage*, 32(1):228–237, 2006.
- [19] PK McGuire and CD Frith. Disordered functional connectivity in schizophrenia. *Psychological medicine*, 26(04):663–667, 1996.
- [20] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.
- [21] Geoffrey J McLachlan and Kaye E Basford. Mixture models. inference and applications to clustering. *Statistics: Textbooks and Monographs, New York: Dekker, 1988*, 1, 1988.
- [22] Sifis Micheloyannis, Ellie Pachou, Cornelis Jan Stam, Michael Breakspear, Panagiotis Bitsios, Michael Vourkas, Sophia Erimaki, and Michael Zervakis. Small-world networks and disturbed functional connectivity in schizophrenia. *Schizophrenia research*, 87(1): 60–66, 2006.
- [23] William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical Parametric Mapping: The Analysis of Functional Brain Images: The Analysis of Functional Brain Images*. Academic Press, 2011.
- [24] Russell A Poldrack. Region of interest analysis for fmri. *Social cognitive and affective neuroscience*, 2(1):67–70, 2007.
- [25] Huitong Qiu, Fang Han, Han Liu, and Brian Caffo. Joint estimation of multiple graphical models from high dimensional time series. *arXiv preprint arXiv:1311.0219*, 2013.

- [26] Baxter P Rogers, Victoria L Morgan, Allen T Newton, and John C Gore. Assessing functional connectivity in the human brain by fmri. *Magnetic resonance imaging*, 25(10):1347–1357, 2007.
- [27] CB Saper and TC Chelimsky. A cytoarchitectonic and histochemical study of nucleus basalis and associated cell groups in the normal human brain. *Neuroscience*, 13(4):1023–1037, 1984.
- [28] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.
- [29] Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.
- [30] Padhraic Smyth. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, 10(1):63–72, 2000.
- [31] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [32] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.
- [33] Jean Esther Toncray and Wendell JS Krieg. The nuclei of the human thalamus: a comparative approach. *Journal of Comparative Neurology*, 85(3):421–459, 1946.
- [34] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

- [35] Gaël Varoquaux and R Cameron Craddock. Learning and comparing functional connectomes across subjects. *NeuroImage*, 80:405–415, 2013.
- [36] Kun Wang, Meng Liang, Liang Wang, Lixia Tian, Xinqing Zhang, Kuncheng Li, and Tianzi Jiang. Altered functional connectivity in early alzheimer’s disease: A resting-state fmri study. *Human brain mapping*, 28(10):967–978, 2007.

Conclusion

In this dissertation, we presented three innovative research projects which make contributions to improved tuning of imaging systems, generalized features for image classification, and a model for functional connectivity in molecular imaging. Each project has made a unique contribution to the field of molecular neuroimaging using methods from machine learning and probabilistic graphical modeling. Each of the projects stands on its own, yet when considered together, improves the overall workflow from data collection to analysis in molecular imaging. Each of these projects were conceived to solve existing problems or improve methods in research directions of interest to investigators at the University of California, Irvine's Neuroscience Imaging Center. Here we briefly describe the future directions for each of these projects, beyond those highlighted in each of the chapters, with a focus on making them accessible to the broader research and clinical communities.

5.1 Future Directions

In chapter 2 we developed a model to improve tuning of PET imaging systems. In order for other sites to use this improved model, the code must be incorporated into the existing manufacturers tuning workflow. Based on discussions with the HRRT PET community and Siemens representatives, we have developed a plan to incorporate our model as a post pro-

cessing step in the typical camera tuning workflow, prior to the typical steps performed by local site engineers. By incorporating our model into the workflow, engineers will benefit from improved detector setups and spend less time and resources manually fixing errors made by the current tuning workflow. Next, we will work on applying the model to newer PET platforms with different detector panel designs. Our model should be readily applicable to a variety of panel designs.

In chapter 3 we introduced a novel technique for defining features and extended it for molecular imaging, where the features were applied to different disorders and molecular imaging platforms (i.e. SPECT, PET) with surprisingly accurate results, competitive with disease-specific and modality-specific features tuned for the specific classification problem at hand. In future work we will incorporate the feature pipeline into the clinical practice at the Amen Clinics Inc. and evaluate its performance in separating large cohorts of subjects with a variety of co-morbidities into distinct classes. Dr. Amen has requested I help to develop a classification system that could be used by Amen clinic physician's to better incorporate brain imaging data into their subject-specific personalized treatment plans. Currently the Amen clinic physician visually inspect the brain scans and development treatment plans based on their educated "reading" of brain imaging data, along with other clinical and behavioral data. A classification system will give additional information and quantitative metrics to help in determining co-morbidities (based on probabilities for the various classes) and/or modifying proposed treatment plans based on how close a new brain imaging dataset is to prior subjects scanned at the clinic that may have responded better to one treatment or the other.

Lastly, in chapter 4 we present a model for group-based functional connectivity in molecular imaging and provide a quantitative analysis of how well different sample sizes performed in recovering a ground truth profile. Evaluating functional connectivity in molecular imaging

is a missing component to current group-based image analysis. In order to help investigators use this technique in their data, a clean, production ready code base is needed and should be made available in typical neuroimaging repositories such as the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC). Because the sparse inverse covariance methods used in this research are currently implemented and available in the R statistical package (<http://www.r-project.org/>), porting our code from Matlab to R is the obvious choice to facilitate more widespread use of the method. Next, we will incorporate the Neuroimaging Data Model (NIDM;<http://nidm.nidash.org/>) into the code base to keep track of model parameters, choices made by the user, and overall provenance in a structured and semantically meaningful framework [1]. NIDM is rapidly becoming integrated into popular neuroimaging analysis software and has the capabilities and support to change the way neuroimaging metadata is shared. I have been leading the working group responsible for the creation of the NIDM standard which is based on prior research I did in the Function Biomedical Informatics Research Network (FBIRN) . Incorporating NIDM into the code bases of the research presented in this dissertation and continuing the adoption of NIDM as an international standard for sharing metadata in neuroimaging is another future research direction. Lastly, functional connectivity as assessed through sparse inverse covariance estimation has been used in studying resting-state functional magnetic resonance imaging data yet no quantitative data exists on how well these models perform as a function of sample sizes. Currently researchers are reporting the results with little guidance on how well these models recover the true underlying functional connectivity. Extending our current methodology on group-based static molecular imaging to time-dependent functional magnetic resonance imaging will contribute additional knowledge on how sparse models perform in these settings and how much to trust the results as a function of sample sizes.

In conclusion, I look forward to a career integrating machine learning methods into mainstream neuroimaging and biomedical informatics, an increasingly popular field becoming a

regular topic at conferences. With the continued improvement in models from the machine learning community and cross-disciplinary researchers who do research in both domains, we expect the techniques will make a dramatic impact on our understanding of the brain and, with any luck, aid clinicians in improving patient care.

Bibliography

- [1] David B Keator, K Helmer, Jason Steffener, Jessica A Turner, Theo GM Van Erp, Syam Gadde, N Ashish, GA Burns, and B Nolan Nichols. Towards structured sharing of raw and derived neuroimaging data across existing resources. *Neuroimage*, 82:647–661, 2013.

Appendices

Inverse covariance model results without clustering

The results showing the average area under the TP/FP edge curves (AUC) across the subsets, by sample sizes, using the gold standard for parameter settings as described in section 4.3.0.1 but without the clustering model are shown in Table A.1. The functional data was averaged in each region of interest instead of using the results from the Gaussian mixture model described in section 4.2.1. The AUCs are statistically lower in the GL model (i.e. negative t-scores) than both the FGL and GGL models at data set sizes from 50-2000, with the moderate sample sizes showing the largest decreases. These results are consistent with those when using the clustering model, yet the differences between inverse covariance models are lower in magnitude and the AUCs are slightly higher, likely due to the smaller number of variables relative to sample size.

The results showing the average minimum SSEs across the subsets, by sample sizes, using the gold standard for parameter settings as described in section 4.3.0.1 but without the clustering model are shown in Table A.2. Again, the results are consistent with those presented using the clustering model, but the SSEs are less accurate over the range of sample sizes likely related to averaging functional signals with noise in the larger regions of interest.

Table A.1: TP/FP Edge Area Under Curve (AUC) T-test comparison by model and sample size using gold standard for parameter settings and simple averaging of functional signal in AAL regions of interest

	GL	FGL	GGL	GL-FGL	GL-GGL	FGL-GGL
N	$\overline{AUC} \pm std$	$\overline{AUC} \pm std$	$\overline{AUC} \pm std$	T(P Bonferroni)	T(P Bonferroni)	T(P Bonferroni)
2500	0.840±0.016	0.851±0.016	0.856±0.011	-1.538(1.00E+00)	-2.603(7.01E-01)	0.814(1.00E+00)
2000	0.823±0.009	0.842±0.011	0.842±0.010	-4.157(2.31E-02)	-4.388(1.38E-02)	-0.043(1.00E+00)
1500	0.806±0.011	0.828±0.010	0.824±0.007	-4.517(1.04E-02)	-4.421(1.29E-02)	-0.773(1.00E+00)
1000	0.772±0.009	0.801±0.011	0.796±0.009	-6.398(1.96E-04)	-5.937(5.00E-04)	-1.184(1.00E+00)
750	0.751±0.009	0.782±0.009	0.775±0.007	-7.511(2.32E-05)	-6.623(1.26E-04)	-1.933(1.00E+00)
500	0.720±0.011	0.753±0.012	0.742±0.009	-6.426(1.86E-04)	-5.053(3.23E-03)	-2.192(1.00E+00)
250	0.656±0.011	0.682±0.014	0.675±0.013	-4.600(8.67E-03)	-3.505(9.85E-02)	-1.195(1.00E+00)
100	0.617±0.011	0.638±0.011	0.631±0.009	-4.352(1.50E-02)	-3.177(2.04E-01)	-1.531(1.00E+00)
50	0.576±0.014	0.604±0.013	0.597±0.011	-4.502(1.07E-02)	-3.708(6.27E-02)	-1.189(1.00E+00)
25	0.560±0.014	0.576±0.011	0.574±0.008	-2.869(3.98E-01)	-2.622(6.75E-01)	-0.591(1.00E+00)
10	0.547±0.013	0.559±0.015	0.552±0.010	-1.878(1.00E+00)	-0.945(1.00E+00)	-1.199(1.00E+00)

Table A.2: Minimum average sum of squared errors (SSE) by model and sample size using gold standard for parameter settings and simple averaging of functional signal in AAL regions of interest

	GL	FGL	GGL	GL-FGL	GL-GGL	FGL-GGL
N	$\overline{SSE} \pm std$	$\overline{SSE} \pm std$	$\overline{SSE} \pm std$	T(P Bonferroni)	T(P Bonferroni)	T(P Bonferroni)
2500	5.94±5.08	7.13±6.11	9.59±8.30	-0.472(1.00E+00)	-1.184(1.00E+00)	-0.753(1.00E+00)
2000	7.69±8.88	9.07±7.79	14.36±12.14	0.397(1.00E+00)	-0.800(1.00E+00)	-1.159(1.00E+00)
1000	17.39±14.67	11.03±9.49	22.79±19.20	1.150(1.00E+00)	-0.707(1.00E+00)	-1.736(1.00E+00)
750	21.90±18.63	13.19±11.34	25.88±22.59	1.263(1.00E+00)	-0.430(1.00E+00)	-1.588(1.00E+00)
500	33.68±28.56	19.42±17.05	23.84±20.19	1.356(1.00E+00)	0.890(1.00E+00)	-0.529(1.00E+00)
250	44.23±37.35	42.54±37.06	41.11±34.53	0.102(1.00E+00)	0.194(1.00E+00)	0.089(1.00E+00)
100	78.07±67.43	60.83±51.49	69.76±60.06	0.643(1.00E+00)	0.291(1.00E+00)	-0.357(1.00E+00)
50	116.36±98.85	94.76±79.63	101.89±85.90	0.538(1.00E+00)	0.350(1.00E+00)	-0.192(1.00E+00)
25	188.52±159.39	165.43±140.50	180.41±153.59	0.344(1.00E+00)	0.116(1.00E+00)	-0.228(1.00E+00)
10	241.64±204.92	224.41±189.45	231.57±195.92	0.195(1.00E+00)	0.112(1.00E+00)	-0.083(1.00E+00)