M. Jorge Cardoso
Ivor Simpson
Tal Arbel
Doina Precup
Annemie Ribbens (Eds.)

LNCS 8677

Bayesian and grAphical Models for Biomedical Imaging

First International Workshop, BAMBI 2014
Cambridge, MA, USA, September 18, 2014
Revised Selected Papers

Springer

Lecture Notes in Computer Science 8677

M. Jorge Cardoso Ivor Simpson Tal Arbel
Doina Precup Annemie Ribbens (Eds.)

Bayesian and grAphical Models for Biomedical Imaging

First International Workshop, BAMBI 2014
Cambridge, MA, USA, September 18, 2014
Revised Selected Papers

Springer

Volume Editors

M. Jorge Cardoso
Ivor Simpson
University College London, Centre for Medical Imaging
Front Engineering Building, Malet Place, London WC1E 6BT, UK
E-mail: {m.jorge.cardoso, ivor.simpson}@ucl.ac.uk

Tal Arbel
McGill University, Centre for Intelligent Machines
McConnell Engineering Building, 3480 University Street
Montreal, QC, H3A 0E9, Canada
E-mail: arbel@cim.mcgill.ca

Doina Precup
McGill University, School of Computer Science
McConnell Engineering Building, 3480 University Street
Montreal, QC, H3A 0E9, Canada
E-mail: dprecup@cs.mcgill.ca

Annemie Ribbens
Katholieke Universiteit Leuven, Medical Imaging Research Center
UZ Herestraat 49, Box 7003, 3000 Leuven, Belgium
E-mail: annemie.ribbens@uzleuven.be

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-319-12288-5 e-ISBN 978-3-319-12289-2
DOI 10.1007/978-3-319-12289-2
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014950465

LNCS Sublibrary: SL 1 – Theoretical Computer Science and General Issues

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

BAMBI 2014 was the First International Workshop on Bayesian and grAphical Models for Biomedical Imaging. It was held at the MIT/Harvard Medical School, Cambridge, MA, USA, on September 18, 2014. This goal of this event was to highlight the potential of using Bayesian or random field graphical models for advancing scientific research in biomedical image analysis.

The BAMBI 2014 proceedings published in the *Lecture Notes in Computer Science* series contain state-of-the-art original and highly methodological research selected through a rigorous peer-review process. Every full paper (10 to 12 pages long in the proceedings format) went through a double-blind review process by at least three members of the international Program Committee composed of 21 renowned scientists in the field of Bayesian image analysis. The result of this selection process was a set of 11 articles, nine of which were selected for oral presentation, and all of which were presented as posters, in a single-track single-day event.

The scientific program was augmented by our three invited speakers, Koen Van Leemput (Athinoula A. Martinos Center for Biomedical Imaging Massachusetts General Hospital, Harvard Medical School, USA and the Department of Applied Mathematics and Computer Science, Technical University of Denmark), Mike Miller (Center for Imaging Science, John Hopkins University, USA), and Ramin Zabih (Cornell University, USA). All three presented exciting advances during their keynote lectures, covering a large scope of methodologies and applications in Bayesian and graphical models.

We warmly thank the members of our Program Committee and all the participants of the event who made this workshop an exciting venue to share the latest methodological advances in this expanding research area.

September 2014

M. Jorge Cardoso
Ivor Simpson
Tal Arbel
Doina Precup
Annemie Ribbens

Organization

The First International Workshop on Bayesian and grAphical Models for Biomedical Imaging (BAMBI 2014) was organized by:

Organizing Committee

M. Jorge Cardoso	Centre for Medical Image Computing, University College London, UK
Ivor Simpson	Centre for Medical Image Computing, University College London, UK
Tal Arbel	Centre for Intelligent Machines, McGill University, Montreal, Canada
Doina Precup	School of Computer Science, McGill University, Montreal, Canada
Annemie Ribbens	Medical Imaging Research Center, KU Leuven and IcoMetrix, Leuven, Belgium

Advisory Panel

Nikos Paragios	Center for Visual Computing, Ecole Centrale de Paris, Paris, France
Koen van Leemput	Harvard Medical School, USA, Technical University of Denmark, Denmark
John Ashburner	Wellcome Trust Centre for Neuroimaging, University College London, UK
William M. Wells III	Harvard Medical School and Brigham and Women's Hospital, Boston, USA

Program Committee

Albert Chung	Hong Kong University of Science and Technology, SAR China
Andre Marquand	King's College London, UK
Anuj Srivastava	Florida State University, USA
Bennett Landeman	Vanderbilt University, USA
Carole Sudre	University College London, UK
Tom Fletcher	University of Utah, USA
Frederik Maes	KU Leuven, Belgium
Ged Ridgway	Oxford University, UK

James Gee	University of Pennsylvania, USA
Ghassan Hamarmeh	Simon Fraser University, Canada
Herve Lombaert	Microsoft Research/Inria, France
John Aston	Cambridge University, UK
Johan Nuyts	KU Leuven, Belgium
Killian Pohl	SRI International, USA
Mattias Heinrich	University of Lübeck, Germany
Paulo Mendonca	GE Global Research, Niskayuna, USA
Mert Sabuncu	Harvard Medical School/MIT, USA
Pankaj Daga	University College London, UK
Stefano Pedemonte	Harvard Medical School/MIT, USA
Suyash Awate	Indian Institute of Technology, India
Yipeng Hu	University College London, UK

Table of Contents

N3 Bias Field Correction Explained as a Bayesian Modeling Method

Christian Thode Larsen[1], J. Eugenio Iglesias[2,3], and Koen Van Leemput[1,2,4]

[1] Department of Applied Mathematics and Computer Science,
Technical University of Denmark
[2] Martinos Center for Biomedical Imaging, MGH, Harvard Medical School, USA
[3] Basque Center on Cognition, Brain and Language, Spain
[4] Departments of Information and Computer Science and of Biomedical Engineering
and Computational Science, Aalto University, Finland

Abstract. Although N3 is perhaps the most widely used method for MRI bias field correction, its underlying mechanism is in fact not well understood. Specifically, the method relies on a relatively heuristic recipe of alternating iterative steps that does not optimize any particular objective function. In this paper we explain the successful bias field correction properties of N3 by showing that it implicitly uses the same generative models and computational strategies as expectation maximization (EM) based bias field correction methods. We demonstrate experimentally that purely EM-based methods are capable of producing bias field correction results comparable to those of N3 in less computation time.

1 Introduction

Due to its superior image contrast in soft tissue without involving ionizing radiation, magnetic resonance imaging (MRI) is the *de facto* modality in brain studies, and it is widely used to examine other anatomical regions as well. MRI suffers from an imaging artifact commonly referred to as "intensity inhomogeneity" or "bias field", which appears as low-frequency multiplicative noise in the images. This artifact is present at all magnetic field strengths, but is more prominent at the higher fields that see increasing use (e.g., 3T or 7T data). Since intensity inhomogeneity negatively impacts any computerized analysis of the MRI data, its correction is often one of the first steps in MRI analysis pipelines.

A number of works have proposed bias field correction methods that are integrated into tissue classification algorithms, typically within the domain of brain MRI analysis [1–7]. These methods often rely on generative probabilistic models, and combine Gaussian mixtures to model the image intensities with a spatially smooth, multiplicative model of the bias field artifact. Cast as a Bayesian inference problem, fitting these models to the MRI data employs expectation-maximization (EM) [8] optimizers to estimate some [7] or all [1, 3, 4, 6] of the model parameters. Specifically tailored for brain MRI analysis applications, these methods encode strong prior knowledge about the number and spatial distribution of tissue types present in the images. As such, they cannot be used out of the box to bias field correct imaging data from arbitrary anatomical regions.

M.J. Cardoso et al. (Eds.): BAMBI 2014, LNCS 8677, pp. 1–12, 2014.

In contrast, the popular N3 [9] bias field correction algorithm does not require any prior information about the MRI input. This allows N3 to correct images of various locations and contrasts, and even automatically handle images that contain pathology. However, despite excellent performance and widespread use, its underlying bias field correction mechanism is not well understood. Specifically, the original paper [9] presents N3 as a relatively heuristic recipe for increasing the "frequency content" of the histogram of an image, by performing specific iterative steps without optimization of any particular objective function.

This paper aims to demonstrate how N3 is in fact intimately linked to EM-based bias field correction methods. In particular, N3 uses the same generative models and bias field estimation computations; however, instead of using dedicated Gaussian mixture models that encode specific prior anatomical knowledge, N3 uses generic models with a very large number of components (200) that are fitted to the histogram by a regularized least-squares method.

The contribution of this paper is twofold. First, to the best of our knowledge, this is the first study offering theoretical insight into why the seemingly heuristic N3 iterations yield such successful bias field estimations. Second, we demonstrate experimentally on datasets of 3T and 7T brain scans that standard EM-based methods, using far less components, are able to produce comparable bias field estimation performance at reduced computational cost.

2 Methods

In this section, we first describe the N3 bias field correction method and its practical implementation. We then present EM-based bias field correction and the generative model it is based upon. Finally, we build an analogy between the two methods, thereby pointing out their close similarities.

2.1 The N3 Method in Its Practical Implementation

The following description is based on version 1.12[1] of the N3 method. In order to facilitate relating the method to a generative model in subsequent sections, we deviate from the notational conventions used in the original paper [9]. Furthermore, whereas the original paper only provides a high-level description of the algorithm (including integrals in the continuous domain, etc.), here we describe the actual implementation in which various discretization, interpolation, and other processing steps are performed.

Let $\boldsymbol{d} = (d_1, \ldots, d_N)^T$ be the intensities of the N voxels of a MRI scan, and let $\boldsymbol{b} = (b_1, \ldots, b_N)^T$ be the corresponding gains due to the bias field. As commonly done in the bias field correction literature [1, 3, 4, 6], N3 assumes that $\boldsymbol{d}$ and $\boldsymbol{b}$ have been log-transformed, such that the effect of $\boldsymbol{b}$ is additive. The central idea behind N3 is that the histogram of $\boldsymbol{d}$ is a blurred version of the histogram of the true, underlying image due to convolution with the histogram of $\boldsymbol{b}$, under the

[1] Source code freely available from `http://packages.bic.mni.mcgill.ca/tgz/`.

assumption that $\boldsymbol{b}$ has the shape of a zero-mean Gaussian with known variance. The algorithm aims to reverse this by means of Wiener deconvolution and to estimate a smooth bias field model accordingly. This reversal process is repeated iteratively, because it was found to improve the bias field estimates [9].

Deconvolution Step: The first step of the algorithm is to deconvolve the histogram. Given the current bias field estimate denoted $\tilde{\boldsymbol{b}}$, a normalized histogram with $K = 200$ bins of bias field corrected data $\boldsymbol{d} - \tilde{\boldsymbol{b}}$ is computed[2]. The bin centers are given by

$$\tilde{\mu}_1 = \min(\boldsymbol{d} - \tilde{\boldsymbol{b}}), \quad \tilde{\mu}_K = \max(\boldsymbol{d} - \tilde{\boldsymbol{b}}), \quad \tilde{\mu}_k = \tilde{\mu}_1 + (k-1)h, \tag{1}$$

where $h = (\tilde{\mu}_K - \tilde{\mu}_1)/(K-1)$ is the bin width, and the histogram entries $\{v_k, k = 1, \ldots, K\}$ are filled using the following interpolation model:

$$v_k = \frac{1}{N}\sum_{i=1}^{N} \varphi\left[\frac{d_i - \tilde{b}_i - \tilde{\mu}_k}{h}\right], \qquad \varphi[s] = \begin{cases} 1 - |s| & \text{if } |s| < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Defining $\hat{\boldsymbol{v}}$ as a padded, 512-dimensional vector such that $\hat{\boldsymbol{v}} = (\mathbf{0}_{156}^T, \boldsymbol{v}^T, \mathbf{0}_{156}^T)^T$, where $\boldsymbol{v} = (v_1, \ldots, v_K)^T$ and $\mathbf{0}_{156}$ is an all-zero 156-dimensional vector, the histogram is deconvolved by

$$\hat{\boldsymbol{\pi}} \leftarrow \boldsymbol{F}^{-1}\boldsymbol{D}\boldsymbol{F}\hat{\boldsymbol{v}}. \tag{2}$$

Here $\boldsymbol{F}$ denotes the 512 × 512 Discrete Fourier Transform matrix with elements

$$F_{n,k} = e^{-2\pi j(k-1)(n-1)/512}, \quad n, k = 1, \ldots, 512$$

and $\boldsymbol{D}$ is a 512 × 512 diagonal matrix with elements

$$D_k = \frac{f_k^*}{|f_k|^2 + \gamma}, \quad k = 1, \ldots, 512$$

where γ is a constant value set to $\gamma = 0.1$, and $\boldsymbol{f} = (f1, \ldots, f_{512})^T = \boldsymbol{F}\boldsymbol{g}$. Here $\boldsymbol{g}$ denotes a 512-dimensional vector that contains a wrapped Gaussian kernel with variance

$$\tilde{\sigma}^2 = \frac{f^2}{8\log 2}, \tag{3}$$

such that

$$\boldsymbol{g} = (g_1, \ldots, g_{512})^T, \quad g_l = \begin{cases} h\mathcal{N}((l-1)h|0, \tilde{\sigma}^2) & \text{if } l = 1, \ldots, 256 \\ g_{512-l+1}, & \text{otherwise,} \end{cases} \tag{4}$$

where f denotes a user-specified full-width-at-half-maximum parameter (0.15 by default), and $\mathcal{N}(\cdot|\mu, \sigma^2)$ denotes a Gaussian distribution with mean μ and variance σ^2.

After $\hat{\boldsymbol{\pi}}$ has been computed by means of Eq. (2), any negative weights are set to zero, and the padding is removed in order to obtain the central deconvolved 200-entry histogram $\tilde{\boldsymbol{\pi}}$.

[2] A flat bias field: $\tilde{\boldsymbol{b}} = \mathbf{0}$ is assumed in the first iteration.

Bias Correction Step: When the histogram $\tilde{\boldsymbol{\pi}}$ has been deconvolved, the corresponding "corrected" intensity $\tilde{d}_{\mu_l}$ in the deconvolved histogram is estimated at each bin center $\tilde{\mu}_l, l = 1, \ldots, K$ by

$$\tilde{d}_{\mu_l} = \sum_k w_k^l \tilde{\mu}_k \quad \text{with} \quad w_k^l = \frac{\mathcal{N}\left(\tilde{\mu}_l | \tilde{\mu}_k, \tilde{\sigma}_k^2\right) \tilde{\pi}_k}{\sum_{k'} \mathcal{N}\left(\tilde{\mu}_l | \tilde{\mu}_{k'}, \tilde{\sigma}_{k'}^2\right) \tilde{\pi}_{k'}},$$

and a "corrected" intensity $\tilde{d}_i$ is found in every voxel by linear interpolation:

$$\tilde{d}_i = \sum_{l=1}^{K} \tilde{d}_{\mu_l} \varphi \left[\frac{d_i - \tilde{b}_i - \tilde{\mu}_l}{h} \right], \qquad \varphi[s] = \begin{cases} 1 - |s| & \text{if } |s| < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Finally, a residual $\boldsymbol{r} = \boldsymbol{d} - \tilde{\boldsymbol{d}}$ is computed and smoothed in order to obtain a bias field estimate:

$$\tilde{\boldsymbol{b}} = \boldsymbol{\Phi} \tilde{\boldsymbol{c}} \tag{5}$$

where

$$\tilde{\boldsymbol{c}} \leftarrow \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + N \beta \boldsymbol{\Psi} \right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{r}. \tag{6}$$

Here $\boldsymbol{\Phi}$ is a $N \times M$ matrix of M spatially smooth basis functions, where element $\Phi_{i,m}$ evaluates the m-th basis function in voxel i; $\boldsymbol{\Psi}$ is a positive semi-definite matrix that penalizes curvature of the bias field; and β is a user-determined regularization constant (the default is $\beta = 10^{-7}$).

Post-Processing: N3 alternates between the deconvolution step and the bias field correction step until the standard deviation of the difference in bias estimates between two iterations drops below a certain threshold (default: $\varsigma = 10^{-3}$). By default, N3 operates on a subsampled volume (factor 4). After convergence, the bias field estimate is exponentiated back into the original intensity domain, where it is subsequently fitted with Eq. (6), i.e., with $\boldsymbol{r} = \exp(\tilde{\boldsymbol{b}})$. The resulting coefficients are then used to compute a final bias field estimate by evaluation of Eq. (5) with $\boldsymbol{\Phi}$ at full image resolution. The uncorrected data is finally divided by the bias field estimate in order to obtain the corrected volume.

2.2 EM-Based Bias Field Estimation

In the following we describe the generative model and parameter optimization strategy underlying EM-based bias field correction methods[3].

[3] Several well-known variants only estimate a subset of the parameters considered here – e.g., in [1] the mixture model parameters are assumed to be known, while [3] uses fixed, spatially varying prior probabilities of tissue types.

Generative Model: Maintaining the notation $\boldsymbol{d}$ to denote a log-transformed image and $\boldsymbol{b} = \boldsymbol{\Phi c}$ to denote a parametric bias field model with parameters $\boldsymbol{c}$, the "true", underlying image $\boldsymbol{d} - \boldsymbol{b}$ is assumed to be a set of N independent samples from a Gaussian mixture model with K components – each with its own mean μ_k, variance σ_k^2, and relative frequency π_k (where $\pi_k \geq 0, \forall k$ and $\sum_k \pi_k = 1$). Given the model parameters $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_k, \sigma_1^2, \ldots, \sigma_K^2, \pi_1, \ldots, \pi_K, c_1, \ldots, c_M)^T$, the probability of an image is therefore

$$p(\boldsymbol{d}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \left[\sum_{k=1}^{K} \mathcal{N}(d_i - \sum_{m=1}^{M} c_m \Phi_{i,m} | \mu_k, \sigma_k^2) \pi_k \right] . \tag{7}$$

The generative model is completed by a prior distribution on its parameters, which is typically of the form

$$p(\boldsymbol{\theta}) \propto \exp[-\lambda \boldsymbol{c}^T \boldsymbol{\Psi} \boldsymbol{c}],$$

where λ is a user-specified regularization hyperparameter and $\boldsymbol{\Psi}$ is a positive semi-definite regularization matrix. This model encompasses approaches where bias field smoothness is imposed either solely through the choice of basis functions (i.e., $\lambda = 0$, as in [3]), or through regularization only (i.e., $\boldsymbol{\Phi} = \boldsymbol{I}$, as in [1]). The prior is uniform with respect to the mixture model parameters.

Parameter Optimization: According to Bayes's rule, the maximum a posteriori (MAP) parameters are given by

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log p(\boldsymbol{\theta}|\boldsymbol{d}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\log p(\boldsymbol{d}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \right] . \tag{8}$$

By exploiting the specific structure of $p(\boldsymbol{d}|\boldsymbol{\theta})$ given by Eq. (7), this optimization can be performed conveniently using a generalized EM (GEM) algorithm [3, 8]. In particular, GEM iteratively builds a lower bound $\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ of the objective function that touches it at the current estimate $\tilde{\boldsymbol{\theta}}$ of the model parameters (E step), and subsequently improves $\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}})$ with respect to the parameters (M step) [8, 10]. This procedure automatically guarantees to increase the value of the objective function at each iteration. Constructing the lower bound involves computing soft assignments of each voxel i to each class k:

$$w_k^i = \frac{\mathcal{N}\left(d_i - \sum_m \tilde{c}_m \Phi_{i,m} | \tilde{\mu}_k, \tilde{\sigma}_k^2\right) \tilde{\pi}_k}{\sum_{k'} \mathcal{N}\left(d_i - \sum_m \tilde{c}_m \Phi_{i,m} | \tilde{\mu}_{k'}, \tilde{\sigma}_{k'}^2\right) \tilde{\pi}_{k'}}, \tag{9}$$

which yields the following lower bound:

$$\varphi(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) = \sum_i \left[\sum_k w_k^i \log \left(\frac{\mathcal{N}(d_i - \sum_m c_m \Phi_{i,m} | \mu_k, \sigma_k^2) \pi_k}{w_k^i} \right) \right] - \lambda \boldsymbol{c}^T \boldsymbol{\Psi} \boldsymbol{c}. \tag{10}$$

Optimizing Eq. (10) simultaneously for the Gaussian mixture model parameters and bias field parameters is difficult. However, optimization with respect to the mixture model parameters for a given set of bias field parameters is closed form:

$$\tilde{\mu}_k \leftarrow \frac{\sum_i w_k^i (d_i - \sum_m \tilde{c}_m \Phi_{i,m})}{\sum_i w_k^i}, \quad \tilde{\sigma}_k^2 \leftarrow \frac{\sum_i w_k^i \left(d_i - \sum_m \tilde{c}_m \Phi_{i,m} - \tilde{\mu}_k\right)^2}{\sum_i w_k^i} \tag{11}$$

$$\tilde{\pi}_k \leftarrow \frac{\sum_i w_k^i}{N}. \tag{12}$$

Similarly, for a given set of mixture model parameters the optimal bias field parameters are given by

$$\tilde{\boldsymbol{c}} \leftarrow \left(\boldsymbol{\Phi}^T \boldsymbol{S} \boldsymbol{\Phi} + 2\lambda \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{S} \boldsymbol{r}, \tag{13}$$

with

$$s_k^i = \frac{w_k^i}{\tilde{\sigma}_k^2}, \quad s_i = \sum_k s_k^i, \quad \boldsymbol{S} = \mathrm{diag}(s_i), \quad \tilde{d}_i = \frac{\sum_k s_k^i \tilde{\mu}_k}{\sum_k s_k^i}, \quad \boldsymbol{r} = \boldsymbol{d} - \tilde{\boldsymbol{d}}.$$

Valid GEM algorithms solving Eq. (8) are now obtained by alternately updating the voxels' class assignments (Eq. (9)), the mixture model parameters (Eqns. (11) and (12)), and the bias field parameters (Eq. (13)), in any order or arrangement.

2.3 N3 as an Approximate MAP Parameter Estimator

Having laid out the details of both N3 and EM-based bias field correction, we are in a position to illustrate parallels between these two methods. In particular, as we describe below, *N3 implicitly uses the same generative model as EM methods* and shares the exact same bias field parameter update (up to numerical discretization aspects). The only difference is that, whereas EM methods fit their Gaussian mixture models by maximum likelihood estimation, N3 does so by regularized least-squares fitting of the mixture model to the histogram entries. Thus, whereas N3 was conceived as iteratively deconvolving Gaussian bias field histograms from the data without optimizing any particular objective function, its successful performance can be readily understood from a standard Bayesian modeling perspective.

Considering the generative model described in Section 2.2, we postulate that N3 uses $K = 200$ Gaussian distributions that are equidistantly spaced between the minimum and maximum intensity, i.e., the parameters $\{\mu_k\}$ are fixed (Eq. (1)). Furthermore, all Gaussians are forced to have an identical variance that is also fixed: $\sigma_k^2 = \tilde{\sigma}^2, \forall k$, where $\tilde{\sigma}^2$ is given by Eq. (3). Thus, the only free parameters in N3 are the relative class frequencies $\pi_k, k = 1, \ldots, K$ and the bias field parameters $\boldsymbol{c}$. We start by analyzing the update equations for $\boldsymbol{c}$.

For the specific scenario where $\sigma_k^2 = \tilde{\sigma}^2, \forall k$, the EM bias field update equation (Eq. (13)) simplifies to

$$\tilde{\boldsymbol{c}} \leftarrow \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi} + 2\tilde{\sigma}^2 \lambda \boldsymbol{\Psi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{r}, \quad \text{with} \quad \tilde{d}_i = \sum_k w_k^i \tilde{\mu}_k, \quad \boldsymbol{r} = \boldsymbol{d} - \tilde{\boldsymbol{d}},$$

where w_k^i is given by Eq. (9). When the hyperparameter λ is set to the value $\lambda = N\beta/2/\tilde{\sigma}^2$ this corresponds directly to the N3 bias field update equation Eq. (6), where the only difference is that N3 explicitly computes $\tilde{d}_{\mu_l}$ for just 200 discrete intensity values and interpolates to obtain $\tilde{d}_i$, instead of computing $\tilde{d}_i$ directly for each individual voxel.

For the remaining parameters $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)^T$, N3 implicitly uses a regularized least-squares fit of the resulting mixture model to the zero-padded normalized histogram $\hat{\boldsymbol{v}}$:

$$\hat{\boldsymbol{\pi}} \leftarrow \underset{\boldsymbol{x}}{\operatorname{argmax}} \|\hat{\boldsymbol{v}} - \boldsymbol{A}\boldsymbol{x}\|^2 + \gamma\|\boldsymbol{x}\|^2, \tag{14}$$

where $\boldsymbol{A}$ is a 512×512 matrix in which each column contains the same Gaussian-shaped basis function, shifted by an offset identical to the column index:

$$\boldsymbol{A} = \begin{pmatrix} g_1 & g_{512} & \cdots & g_2 \\ g_2 & g_1 & \cdots & g_3 \\ \vdots & \vdots & \ddots & \vdots \\ g_{512} & g_{511} & \cdots & g_1 \end{pmatrix},$$

i.e., the first column contains the vector $\boldsymbol{g}$ defined in Eq. (4), and the remaining columns contain cyclic permutations of $\boldsymbol{g}$. To see why Eq. (14) is equivalent to Eq. (2), consider that because $\boldsymbol{A}$ is a circulant matrix, it can be decomposed as

$$\boldsymbol{A} = \boldsymbol{F}^{-1}\boldsymbol{\Lambda}\boldsymbol{F} \quad \text{with} \quad \boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{f}),$$

where $\boldsymbol{F}$ and $\boldsymbol{f}$ were defined in Section 2.1. The solution of Eq. (14) is given by

$$\begin{aligned} \hat{\boldsymbol{\pi}} \leftarrow & \left(\boldsymbol{A}^T\boldsymbol{A} + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{A}^T\hat{\boldsymbol{v}} = \left(\boldsymbol{F}^{-1}\boldsymbol{\Lambda}^H\boldsymbol{F}\boldsymbol{F}^{-1}\boldsymbol{\Lambda}\boldsymbol{F} + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{F}^{-1}\boldsymbol{\Lambda}^H\boldsymbol{F}\hat{\boldsymbol{v}} \\ = & \left(\boldsymbol{F}^{-1}\boldsymbol{\Lambda}^H\boldsymbol{\Lambda}\boldsymbol{F} + \gamma\boldsymbol{F}^{-1}\boldsymbol{F}\right)^{-1}\boldsymbol{F}^{-1}\boldsymbol{\Lambda}^H\boldsymbol{F}\hat{\boldsymbol{v}} = \boldsymbol{F}^{-1}\underbrace{\left(\boldsymbol{\Lambda}^H\boldsymbol{\Lambda} + \gamma\boldsymbol{I}\right)^{-1}\boldsymbol{\Lambda}^H}_{\boldsymbol{D}}\boldsymbol{F}\hat{\boldsymbol{v}}, \end{aligned}$$

where $\boldsymbol{A}^H$ denotes the *Hermitian transpose* of $\boldsymbol{A}$ and where we have used the properties that $\boldsymbol{A}^T = \boldsymbol{A}^H$ and $\boldsymbol{F}^H = 512 \cdot \boldsymbol{F}^{-1}$.

An example of N3's mixture model fitted this way will be shown in Figure 1. The periodic end conditions in $\boldsymbol{A}$ have no practical impact on the histogram fit, as the support of the Gaussian-shaped basis functions is limited, and only the parameters of the 200 central basis functions are retained after fitting. Although this is clearly an *ad hoc* approach, the results are certainly not unreasonable, and N3 thereby maintains a close similarity to purely EM-based bias field correction methods.

3 Experiments

Implementation: In order to experimentally verify our theoretical analysis and quantify the effect of replacing the N3 algorithm of Section 2.1 with the EM algorithm described in Section 2.2 and *vice versa*, we implemented both methods in Matlab. For our implementation of N3, we took care to mimic the original N3 implementation (a Perl script binding together a number of C++ binaries) as faithfully as possible. Specifically, we used identically placed cubic B-spline basis functions $\boldsymbol{\Phi}$, identical regularizer $\boldsymbol{\Psi}$, and the same sub-sampling scheme and parameter settings as in the original method. Our EM implementation shares the same characteristics and preprocessing steps where possible, so that any experimental difference in performance between the two methods is explained by algorithmic rather than technological aspects.

During the course of our experiments, we observed that N3's final basis function fitting operation in the original intensity domain (described in Section 2.1, "Post-processing") actually hurts the performance of the bias field correction. Also, we noticed that N3's default threshold value to detect convergence ($\varsigma = 10^{-3}$) tends to stop the iterations prematurely. To ensure a fair comparison with the EM method, we henceforth report the performance of N3 (Matlab) with the final fitting operation switched off, and with a more conservative threshold value that guarantees full convergence of the method ($\varsigma = 10^{-5}$).

For our EM implementation, we report results for mixture models of $K = 3$, $K = 6$, and $K = 9$ components. We initialize the algorithm with the bias field coefficients set to zero: $\boldsymbol{c} = \boldsymbol{0}$ (no bias field); with equal relative class frequencies: $\pi_k = 1/K, \forall k$; equidistantly placed means given by Eq. (1) and equal variances given by $\sigma_k^2 = ((\max(\boldsymbol{d}) - \min(\boldsymbol{d}))/K)^2, \forall k$. For a given bias field estimate, the algorithm alternates between re-computing $w_k^i, \forall i, k$ (Eq. (9)) and updating the mixture model parameters (Eqns. (11) and (12)), until convergence in the objective function is detected (relative change between iterations $< 10^{-6}$). Subsequently, the bias field is updated (Eq. 13) and the whole process is repeated until global convergence is detected (relative change in the objective function $< 10^{-5}$).

MRI Data and Brain Masking: We tested both bias field correction methods on two separate datasets of T1-weighted brain MR scans. The first dataset was acquired on several 3T Siemens Tim Trio scanners using a multi-echo MPRAGE sequence with a voxel size of $1.2 \times 1.2 \times 1.2$ mm^3. It consists of 38 subjects scanned twice with varying intervals for a total of 76 volumes. The second dataset consists of 17 volumes acquired on a 7T Siemens whole-body MRI scanner using a multi-echo MPRAGE sequence with a voxel size of $0.75 \times 0.75 \times 0.75$ mm^3. Since N3 bias field correction of brain images is known to work well only on scans in which all non-brain tissue has been removed [11], both datasets were skull-stripped using FreeSurfer[4].

Evaluation Metrics: Since the true bias field effect in our MR images is unknown, we compare the two methods using a segmentation-based approach. In particular, we use the coefficient of joint variation [12] in the white and gray matter as an evaluation metric, measured in the original (rather than logarithmic)

[4] `https://surfer.nmr.mgh.harvard.edu/`

domain of image intensities, after bias field correction. This metric is defined as CJV $= \frac{\sigma_1+\sigma_2}{|\mu_1-\mu_2|}$, where (μ_1, σ_1) and (μ_2, σ_2) denote the mean and standard deviation of intensities within the white and the gray matter, respectively. Compared to the coefficient of variation defined as CV $= \sigma_1/\mu_1$, which is also commonly used in the literature [11, 13] and which measures only the intensity variation within the white matter, the CJV additionally takes into account the remaining separation between white and gray matter intensities.

In order to compute the CJV, we used FreeSurfer to obtain automatic white and gray matter segmentations, which we then eroded once in order to limit the influence of boundary voxels, which are typically affected by partial volume effects. We observed that the segmentation performance of FreeSurfer was suboptimal in the 7T data because this software has problems with field strengths above 3T. This problem was ameliorated by bias field correcting the 7T scans with SPM8[5] prior to feeding them to FreeSurfer.

In addition to reporting CJV results for the two methods, we also report their run time on a 64bit CentOS 6.5 Linux PC with 24 gigabytes of RAM, an Intel(R) Xeon(R) E5430 2.66GHz CPU, and with Matlab version R2013b installed. For the sake of completeness, we also include the CJV and run time results for the original N3 software (default parameters, with the exception of the spacing between the B-spline control points – see below).

Stiffness of the Bias Field Model: The stiffness of the B-spline bias field model is determined both by the spacing between the B-spline control points (affecting the number of basis functions in $\boldsymbol{\Phi}$) and the regularization parameter of $\boldsymbol{\Psi}$ that penalizes curvature (β in N3, and λ in the EM method).

As recommended in [13], we used a spacing of 50 mm instead of the N3 default[6], as it is known to be too large for images obtained at higher-field strengths. Finding a common, matching value for the regularization parameter in both methods proved difficult, since we observed that the methods perform best in different ranges. Therefore, for the current study we computed average CJV scores for both methods over a wide range of values. We report results for the setting that worked best for each method and for each dataset separately[7].

4 Results

Figure 1 shows the histogram fit and the bias field estimate of both our N3 implementation and the EM method with $K = 6$ Gaussian components on a representative scan from the 7T dataset. In general, the histogram fit works well for both methods; however for N3 a model mismatch can be seen around the high-intensity tail. This is the result of zeroing negative weights after Wiener filtering.

[5] `http://www.fil.ion.ucl.ac.uk/spm/`

[6] 200 mm, appropriate for the 1.5T data the method was originally developed for.

[7] A more elaborate validation study would determine the optimal values on a separate training dataset; however, this is outside the scope of the current workshop paper.

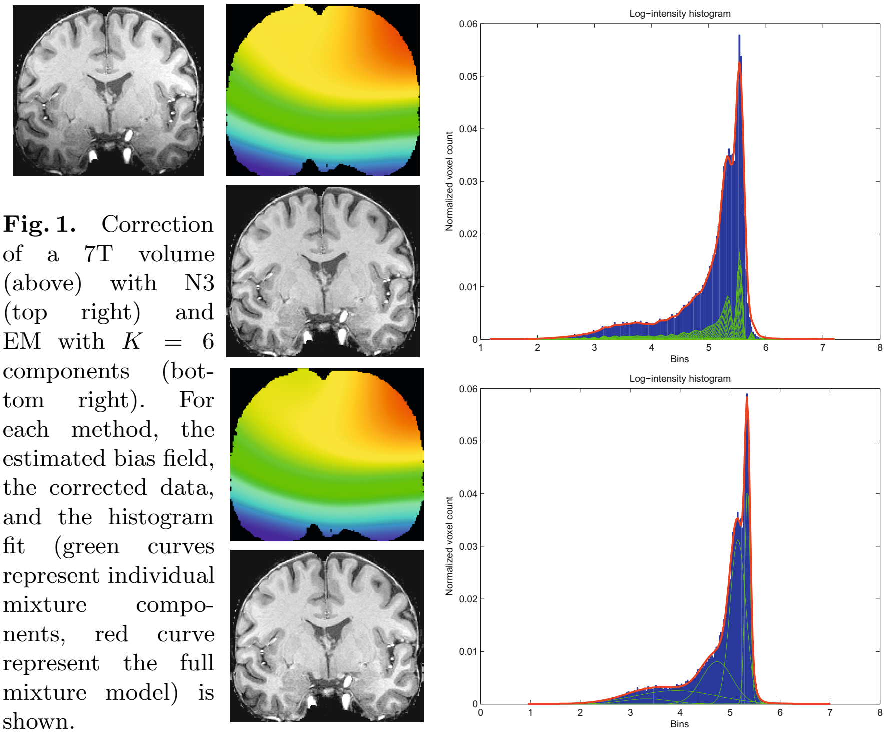


Fig. 1. Correction of a 7T volume (above) with N3 (top right) and EM with $K = 6$ components (bottom right). For each method, the estimated bias field, the corrected data, and the histogram fit (green curves represent individual mixture components, red curve represent the full mixture model) is shown.

Table 1. Average computation time for correcting a volume within each dataset

Dataset	Average computation time (seconds)				
	EM (3G)	EM (6G)	EM (9G)	N3 (Matlab)	N3
3T	12.7	20.7	29.7	86.0	53.5
7T	50.6	79.2	102.0	415.5	170.8

Figure 2 shows the CJV in the two test datasets, before bias field correction as well as after, using the EM method (for $K = 3$, $K = 6$, and $K = 9$ components), our Matlab N3 implementation, and the original N3 software. Overall, the EM and N3 (Matlab) methods perform comparably, except for EM with $K = 3$ components which seems to have too few degrees of freedom in the 7T dataset. The original N3 implementation is provided as a reference only; its underperformance compared to our own implementation is to be expected since its settings were not tuned the same way.

Table 1 shows the average computation time of each method. Due to the much higher resolution of the 7T data, computation time increased for all methods when correcting this dataset. In all cases, the EM correction ran three to six times faster than the N3 Matlab implementation, depending on the number of components in the mixture. As before, results for the original N3 method are provided for reference only.

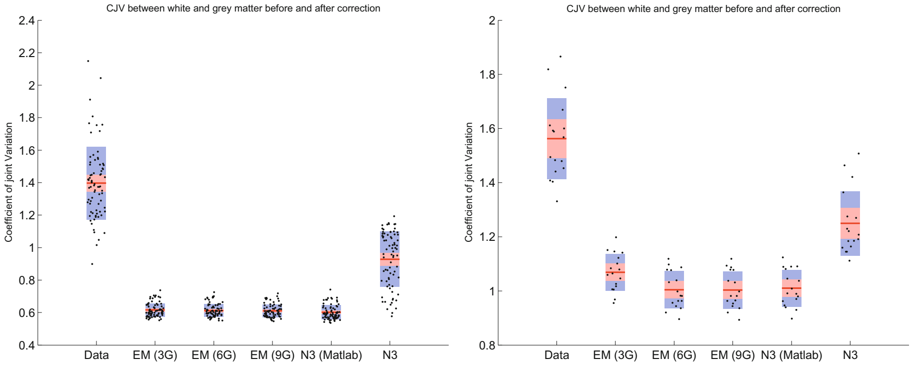


Fig. 2. Scatter plots showing the CVJ between white and gray matter in the 3T (left) and 7T (right) datasets. Lower CVJ equates to better performance. The red line represents the mean, while the blue box covers one standard deviation of the data and the red box covers the 95% confidence interval of the mean.

5 Discussion

In this paper we have explained the successful bias field correction properties of the N3 method by showing that it implicitly uses the same type of generative models and computational strategies as EM-based bias field correction methods. Experiments on MRI scans of healthy brains indicate that, at least in this application, purely EM-based methods can achieve performance similar to N3 at a reduced computational cost.

Future work should evaluate how replacing N3's highly constrained 200-component mixture model with more general mixture models affects bias field correction performance in scans containing pathology. Conversely, while N3's idiosyncratic histogram fitting procedure was found to work well in our experiments, it is worth noting that it precludes N3 from taking advantage of specific prior domain knowledge when such is available. For instance, the skull stripping required to make N3 work well in brain studies [11] typically involves registration of the images into a standard template space, which means that probabilistic brain atlases are available at no additional cost. It is left as further work to evaluate whether this puts N3 at a potential disadvantage compared to EM-based methods, which can easily take this form of extra information into account [3, 7]. Future validation studies should also include comparisons with the publicly available N4ITK implementation [14], which employs a more elaborate but heuristic B-spline fitting procedure in the bias field computations.

Acknowledgments. This research was supported by the NIH NCRR (P41-RR14075), the NIH NIBIB (R01EB013565), TEKES (ComBrain), the Danish Council for Strategic Research (J No. 10-092814) and financial contributions from the Technical University of Denmark. The authors would like to thank Jonathan Polimeni for supplying 7T data for our tests.

References

1. Wells, W.M., Grimson, W.E.L., Kinikis, R., Jolesz, F.A.: Adaptive segmentation of MRI data. IEEE Transactions on Medical Imaging 15(4), 429–442 (1996)
2. Held, K., Kops, E., Krause, B., Wells, W., Kikinis, R., Muller-Gartner, H.: Markov random field segmentation of brain MR images. IEEE Transactions on Medical Imaging 16(6), 878–886 (1997)
3. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based bias field correction of MR images of the brain. IEEE Transactions on Medical Imaging 18(10), 885 (1999)
4. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based tissue classification of MR images of the brain. IEEE Transactions on Medical Imaging 18(10), 897–908 (1999)
5. Pham, D., Prince, J.: Adaptive fuzzy segmentation of magnetic resonance images. IEEE Transactions on Medical Imaging 18(9), 737–752 (1999)
6. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging 20(1), 45–57 (2001)
7. Ashburner, J., Friston, K.J.: Unified segmentation. NeuroImage 26(3), 839–851 (2005)
8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39(1), 1–38 (1977)
9. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging 17(1), 87–97 (1998)
10. Minka, T.P.: Expectation-maximization as lower bound maximization (1998)
11. Boyes, R.G., Gunter, J.L., Frost, C., Janke, A.L., Yeatman, T., Hill, D.L., Bernstein, M.A., Thompson, P.M., Weiner, M.W., Schuff, N., Alexander, G.E., Killiany, R.J., DeCarli, C., Jack, C.R., Fox, N.C.: Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils. NeuroImage 39(4), 1752–1762 (2008)
12. Likar, B., Viergever, M.A., Pernus, F.: Retrospective correction of MR intensity inhomogeneity by information minimization. IEEE Transactions on Medical Imaging 20(12), 1398–1410 (2001)
13. Zheng, W., Chee, M.W., Zagorodnov, V.: Improvement of brain segmentation accuracy by optimizing non-uniformity correction using N3. NeuroImage 48(1), 73–83 (2009)
14. Tustison, N., Avants, B., Cook, P., Zheng, Y., Egan, A., Yushkevich, P., Gee, J.: N4ITK: Improved N3 bias correction. IEEE Transactions on Medical Imaging 29(6), 1310–1320 (2010)

A Bayesian Approach to Distinguishing Interdigitated Muscles in the Tongue from Limited Diffusion Weighted Imaging

Chuyang Ye[1,⋆], Aaron Carass[1], Emi Murano[2],
Maureen Stone[3], and Jerry L. Prince[1]

[1] Department of Electrical and Computer Engineering,
Johns Hopkins University, Baltimore, MD, USA
[2] Department of Radiology and Radiological Sciences,
Johns Hopkins University School of Medicine, Baltimore, MD, USA
[3] Department of Neural and Pain Sciences,
University of Maryland School of Dentistry, Baltimore, MD, USA

Abstract. Fiber tracking in crossing regions is a well known issue in diffusion tensor imaging (DTI). Multi-tensor models have been proposed to cope with the issue. However, in cases where only a limited number of gradient directions can be acquired, for example in the tongue, the multi-tensor models fail to resolve the crossing correctly due to insufficient information. In this work, we address this challenge by using a fixed tensor basis and incorporating prior directional knowledge. Within a maximum a posteriori (MAP) framework, sparsity of the basis and prior directional knowledge are incorporated in the prior distribution, and data fidelity is encoded in the likelihood term. An objective function can then be obtained and solved using a noise-aware weighted ℓ_1-norm minimization. Experiments on a digital phantom and *in vivo* tongue diffusion data demonstrate that the proposed method is able to resolve crossing fibers with limited gradient directions.

Keywords: Diffusion imaging, weighted ℓ_1-norm minimization, prior directional knowledge.

1 Introduction

Diffusion tensor imaging (DTI) provides a noninvasive tool for investigating fiber tracts by imaging the anisotropy of water diffusion [1]. A well known issue in DTI is fiber tracking in crossing regions, where the tensor model is incorrect [2]. Multi-tensor models have been proposed to cope with this issue. For example, [3] and [4] use two-tensor models to recover crossing directions, [5] deconvolves diffusion signals using a set of diffusion basis functions, and [2] uses a sparse reconstruction, where a fixed tensor basis is used to produce the crossing patterns.

[⋆] This work is supported by the Chinese Scholarship Council, NIH/NINDS 1R21NS082891, NIH/NINDS 5R01NS056307, and NIH/NCI 5R01CA133015.

M.J. Cardoso et al. (Eds.): BAMBI 2014, LNCS 8677, pp. 13–24, 2014.

Using the number of gradient directions that is common in clinical research (around 30), these methods are able to resolve crossing fibers.

However, in cases where limited gradient directions are used, current multi-tensor models have insufficient information for successful resolution of crossing fibers. For example, in the tongue, where involuntary swallowing limits the available time for *in vivo* acquisition, usually only a dozen (or so) gradient directions are achievable, and the acquisition usually takes around two or three minutes. Thus, distinguishing interdigitated tongue muscles, which constitute a large percentage of the tongue volume, is very challenging.

In this work, we present a multi-tensor method that incorporates prior directional information within a Bayesian framework to resolve crossing fibers with limited gradient directions. We use a fixed tensor basis and estimate the contribution of each tensor using a maximum a posteriori (MAP) framework. The prior knowledge contains both directional information and a sparsity constraint, and data fidelity is modeled in the likelihood. The resulting objective function can be solved as a noise-aware version of a weighted ℓ_1-norm minimization [6]. The method is evaluated on *in vivo* tongue diffusion images.

2 Methods

2.1 Multi-tensor Model with a Fixed Tensor Basis

Suppose a fixed tensor basis comprises N prolate tensors $\mathbf{D}_i$, whose primary eigenvectors (PEVs) are oriented over the sphere. In this work, $N = 253$, the primary eigenvalue of each basis tensor is equal to 2×10^{-3} mm^2/s, and the second and third eigenvalues are equal to 0.5×10^{-3} mm^2/s. At each voxel, the diffusion weighted signals are modeled as a mixture of the attenuated signals from these tensors. Using the Stejskal-Tanner tensor formulation [7], we have [2]

$$S_k = S_0 \sum_{i=1}^{N} f_i e^{-b\boldsymbol{g}_k^T \mathbf{D}_i \boldsymbol{g}_k} + n_k, \tag{1}$$

where b is the b-value, $\boldsymbol{g}_k$ is the k-th gradient direction, S_0 is the baseline signal without diffusion weighting, f_i is the (unknown) nonnegative mixture fraction for $\mathbf{D}_i$, and n_k is noise. Each $\mathbf{D}_i$ represents a fiber direction given by its PEV. Note that here we do not require $\sum_i f_i = 1$ as in [2]. Assuming K gradient directions are used, by defining $y_k = S_k/S_0$ and $\eta_k = n_k/S_0$, (1) can be written as

$$\boldsymbol{y} = \mathbf{G}\boldsymbol{f} + \boldsymbol{\eta}, \tag{2}$$

where $\boldsymbol{y} = (y_1, y_2, ..., y_K)^T$, $\mathbf{G}$ is a $K \times N$ matrix comprising the attenuation terms $G_{ki} = e^{-b\boldsymbol{g}_k^T \mathbf{D}_i \boldsymbol{g}_k}$, $\boldsymbol{f} = (f_1, f_2, ..., f_N)^T$, and $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_K)^T$.

2.2 Mixture Fraction Estimation with Prior Knowledge

We use MAP estimation to estimate the mixture fractions $\boldsymbol{f}$. Accordingly, we seek to maximize the posterior probability of $\boldsymbol{f}$ given the observations $\boldsymbol{y}$:

$$p(\boldsymbol{f}|\boldsymbol{y}) = \frac{p(\boldsymbol{f})p(\boldsymbol{y}|\boldsymbol{f})}{\int p(\boldsymbol{f})p(\boldsymbol{y}|\boldsymbol{f})\mathrm{d}\boldsymbol{f}} \propto p(\boldsymbol{f})p(\boldsymbol{y}|\boldsymbol{f}). \tag{3}$$

Since at each voxel the number of fiber directions is expected to be small, we first put a Laplace prior into the prior density $p(\boldsymbol{f})$ to promote sparseness: $p(\boldsymbol{f}) \propto e^{-\lambda||\boldsymbol{f}||_1}$. Sparsity alone is not sufficient prior information when the observations do not include a large number of gradient directions (as in diffusion imaging of the *in vivo* tongue). Therefore, we further supplement the prior knowledge with directional information. For example, the muscles in the tongue have fairly regular organization involving an anterior-posterior (A-P) fanning of the genioglossus and vertical muscles, and a left-right (L-R) crossing of the transverse muscle.

Suppose prior information about likely fiber directions, which we call *prior directions* (PDs), were known at each voxel of the tongue. Let the PDs be represented by the collection of vectors $\{\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_P\}$, where P is the number of the PDs at the voxel. Note that the PDs can vary at different locations, and such information could be provided, for example, by deformable registration of a prior template into the tongue geometry. A similarity vector $\boldsymbol{a}$ can be constructed between the directions represented by the basis tensors and the PDs:

$$\boldsymbol{a} = (\max_m |\boldsymbol{v}_1 \cdot \boldsymbol{w}_m|, \max_m |\boldsymbol{v}_2 \cdot \boldsymbol{w}_m|, ..., \max_m |\boldsymbol{v}_N \cdot \boldsymbol{w}_m|)^T, \tag{4}$$

where $\boldsymbol{v}_i$ is the PEV of the basis tensor $\mathbf{D}_i$. We modify the prior density by incorporating the similarity vector: $p(\boldsymbol{f}) \propto e^{-\lambda||\boldsymbol{f}||_1} e^{\gamma \boldsymbol{a}\cdot\boldsymbol{f}}$. In this way, basis tensors closer to the PDs are made to be more likely *a priori*. Note that $\boldsymbol{w}_m$ and $\boldsymbol{v}_i$ are unit vectors and thus each entry in $\boldsymbol{a}$ is in the interval $[0, 1]$. Since $\boldsymbol{f} \geq \mathbf{0}$,

$$\begin{aligned}\lambda||\boldsymbol{f}||_1 - \gamma\boldsymbol{a}\cdot\boldsymbol{f} = \lambda\mathbf{1}\cdot\boldsymbol{f} - \gamma\boldsymbol{a}\cdot\boldsymbol{f} = \lambda(\mathbf{1} - \frac{\gamma}{\lambda}\boldsymbol{a})\cdot\boldsymbol{f} = \lambda(\mathbf{1} - \alpha\boldsymbol{a})\cdot\boldsymbol{f} \\ = \lambda||\mathbf{C}\boldsymbol{f}||_1,\end{aligned} \tag{5}$$

where $\alpha = \frac{\gamma}{\lambda}$ and $\mathbf{C}$ is a diagonal matrix with $C_{ii} = (1 - \alpha a_i)$. Therefore, $p(\boldsymbol{f})$ has a truncated Laplace density given by

$$p(\boldsymbol{f}) = \frac{1}{Z_\mathrm{p}(\alpha, \lambda)} e^{-\lambda||\mathbf{C}\boldsymbol{f}||_1}, \quad \boldsymbol{f} \geq \mathbf{0}, \tag{6}$$

where $Z_\mathrm{p}(\alpha, \lambda)$ is a constant. We require $0 \leq \alpha < 1$ to ensure that $C_{ii} > 0$.

Suppose the noise $\boldsymbol{\eta}$ in (2) follows a Rician distribution; then it can be approximated by a Gaussian distribution when the signal to noise ratio is above 2:1 [8]. Therefore, we model the likelihood term as a Gaussian density: $p(\boldsymbol{y}|\boldsymbol{f}) \propto e^{-||\mathbf{G}\boldsymbol{f}-\boldsymbol{y}||_2^2/\sigma_\eta^2}$, where σ_η is the noise level normalized by S_0. Then, according to (3), we have the posterior density

$$p(\boldsymbol{f}|\boldsymbol{y}) = \frac{1}{Z(\alpha, \lambda, \sigma_\eta, \mathbf{G})} e^{-(||\mathbf{G}\boldsymbol{f}-\boldsymbol{y}||_2^2/\sigma_\eta^2 + \lambda||\mathbf{C}\boldsymbol{f}||_1)}, \tag{7}$$

where $Z(\alpha, \lambda, \sigma_\eta, \mathbf{G})$ is a normalization constant. The MAP estimate of $\boldsymbol{f}$ is found by maximizing $p(\boldsymbol{f}|\boldsymbol{y})$ or $\ln p(\boldsymbol{f}|\boldsymbol{y})$, which leads to

$$\hat{\boldsymbol{f}} = \underset{\boldsymbol{f} \geq \mathbf{0}}{\arg\min} \frac{1}{\sigma_\eta^2} ||\mathbf{G}\boldsymbol{f} - \boldsymbol{y}||_2^2 + \lambda ||\mathbf{C}\boldsymbol{f}||_1 \tag{8}$$

$$= \underset{\boldsymbol{f} \geq \mathbf{0}}{\arg\min} ||\mathbf{G}\boldsymbol{f} - \boldsymbol{y}||_2^2 + \beta ||\mathbf{C}\boldsymbol{f}||_1, \tag{9}$$

where $\beta = \lambda \sigma_\eta^2$. The problem in (9) is a noise-aware version of a weighted ℓ_1-norm minimization as discussed in [6]. We note that this formulation is equivalent to the CFARI objective function developed in [2] when $\alpha = 0$ (i.e., $\mathbf{C} = \mathbf{I}$). Thus, our approach, developed with an alternative Bayesian perspective, should be considered as a generalization of CFARI.

To solve (9), we use a new variable $\boldsymbol{g} = \mathbf{C}\boldsymbol{f}$. Since $\mathbf{C}$ is diagonal and $C_{ii} > 0$, $\mathbf{C}$ is invertible and therefore $\boldsymbol{f} = \mathbf{C}^{-1}\boldsymbol{g}$. Letting $\tilde{\mathbf{G}} = \mathbf{G}\mathbf{C}^{-1}$, we have

$$\hat{\boldsymbol{g}} = \underset{\boldsymbol{g} \geq \mathbf{0}}{\arg\min} ||\tilde{\mathbf{G}}\boldsymbol{g} - \boldsymbol{y}||_2^2 + \beta ||\boldsymbol{g}||_1. \tag{10}$$

We find $\hat{\boldsymbol{g}}$ using the optimization method in [9] and the mixture fractions $\boldsymbol{f}$ can be estimated as:

$$\hat{\boldsymbol{f}} = \mathbf{C}^{-1}\hat{\boldsymbol{g}}. \tag{11}$$

Directions associated with nonzero mixture fractions are interpreted as fiber directions, and the value of f_i indicates the contribution of the corresponding direction in the diffusion signal. In practice, as in [2], we only keep the directions with the largest 5 mixture fractions f_{n_i} $(i = 1, 2, 3, 4, 5)$ to save memory, which is sufficient to represent all fiber directions. Finally, the mixture fractions are normalized so that they sum to one: $\tilde{f}_{n_i} = f_{n_i} / \sum_{i=1}^{5} f_{n_i}$.

3 Experiments

3.1 Digital Phantom

A 3D crossing phantom with two tracts crossing at 90° was generated to validate the proposed algorithm (see Fig. 1 for an axial view). Twelve gradient directions were used. CFARI [2] and our proposed method were applied on the phantom.

First, we used horizontal and vertical directions as PDs for the horizontal and vertical tracts, respectively. An example of reconstructed directions (for $\alpha = 0.5$ and $\beta = 0.05$) is shown in Fig. 1(b), and is compared with CFARI results in Fig. 1(a). The standard color scheme in DTI is used. Directions with small $\tilde{f}_{n_i}$'s are interpreted as components of isotropic diffusion; therefore we only show directions with $\tilde{f}_{n_i} > 0.1$. It can be seen that in crossing regions, CFARI fails to produce the correct configuration while the proposed method successfully generates the correct crossing pattern.

Next, we studied the effect of inaccurate PDs. To introduce errors in the PDs, we rotated the true directions by $\theta = 15°$ to obtain PDs. We tested two cases

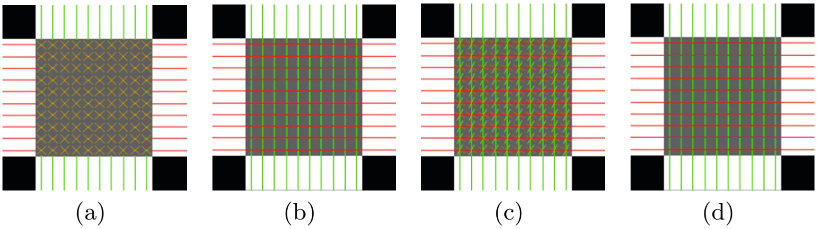


Fig. 1. Axial view of the FA of the crossing phantom. Reconstructed fiber directions are shown for (a) CFARI and (b)–(d) the proposed method. The PDs are ground truth directions in (b), ground truth directions with 15° in-plane rotation in (c), and ground truth directions with 15° out-of-plane rotation in (d).

of rotations: in and out of the axial plane. Specifically, in the first case, the horizontal and vertical directions are both clockwise rotated in the axial plane; and in the second case, the horizontal directions are rotated around the vertical line out of the axial plane and the vertical directions are rotated around the horizontal line out of the axial plane. The results are shown in Figs. 1(c) and 1(d) for the two cases, respectively. In both cases, the proposed method correctly reconstructs noncrossing fiber directions. For the PDs with in-plane rotation, the proposed method is still able to find the crossing directions, although it also produces incorrect fiber directions. For the PDs with out-of-plane rotation, the proposed method successfully reconstructs the crossing directions.

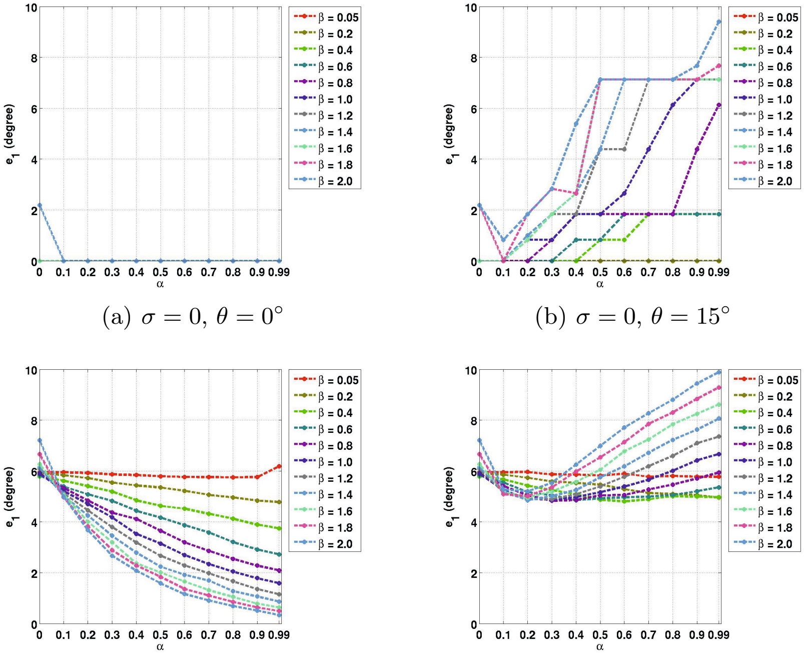
To make the simulation more realistic, besides the noise-free phantom test, Rician noise ($\sigma = 0.06$) was also added to the digital phantom. And we tested with different values of α and β. To quantitatively evaluate the results, we define two error measures:

$$e_1 = \frac{1}{N_1} \sum_{\substack{i=1 \\ \tilde{f}_{n_i} > t}}^{5} \min_j \arccos(\boldsymbol{v}_{n_i} \cdot \boldsymbol{u}_j) \cdot \frac{180^\circ}{\pi} \tag{12}$$

$$e_2 = \frac{1}{N_2} \sum_{j=1}^{N_2} \min_{i:\tilde{f}_{n_i} > t} \arccos(\boldsymbol{v}_{n_i} \cdot \boldsymbol{u}_j) \cdot \frac{180^\circ}{\pi}. \tag{13}$$

Here N_1 is the number of directions with normalized mixture fractions $\tilde{f}_{n_i}$ larger than a threshold t (in this case $t = 0.1$), $\boldsymbol{v}_{n_i}$ is the reconstructed fiber direction, and N_2 is the number of ground truth crossing directions $\boldsymbol{u}_j$. N_2 can be 1 or 2, depending on whether fiber crossing exists at the location. e_1 measures if the reconstructed directions are away from the ground truth, and e_2 measures if each true direction is properly reconstructed. Note that using only e_1 or e_2 is insufficient because the reconstructed directions can agree well with one of the true crossing directions and ignore the other, or each true direction is properly reconstructed but there are other incorrect reconstructed directions.

The average errors in the noncrossing and crossing regions are plotted in Figs. 2 to 5. Here we used the true fiber directions and their 15° rotated versions



(a) $\sigma = 0$, $\theta = 0°$

(b) $\sigma = 0$, $\theta = 15°$

(c) $\sigma = 0.06$, $\theta = 0°$

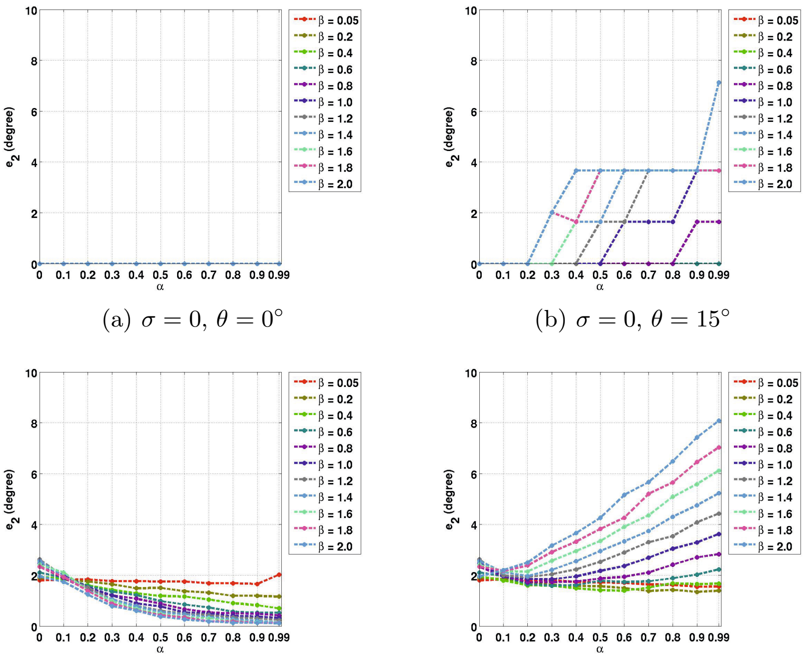
(d) $\sigma = 0.06$, $\theta = 15°$

Fig. 2. Average e_1 errors in noncrossing regions with different noise level σ, PD inaccuracy θ, and the parameters of α and β

as PDs. For the rotated directions, the results in the in-plane and out-of-plane cases are averaged. Note that $\alpha = 0$ is equivalent to CFARI results.

In noncrossing regions, from Figs. 2 and 3, it can be seen that when errors are introduced in the PDs, the correct fiber directions can still be obtained with proper weighting of prior knowledge. For example, as shown in Figs. 2(b) and 3(b), $\alpha = 0.3$ and $\beta = 0.6$ give zero e_1 and e_2 errors. When noise is added, the use of ground truth as PDs leads to more accurate estimation, as shown in Figs. 2(c) and 3(c). When an error of 15° is introduced, the proposed method can still reduce the effect of noise with proper α and β (see $\alpha = 0.5$ and $\beta = 0.6$ in Figs. 2(d) and 3(d)).

In crossing regions, the use of ground truth as PDs produces correct crossing directions in both the noise-free and the noisy cases (see Figs. 4(a), 4(c), 5(a) and 5(c)). When errors are introduced in the PDs, in both the noise-free and the noisy cases, it is still possible to obtain crossing directions that are close to truth with proper α and β (for example, $\alpha = 0.6$ and $\beta = 0.05$ in Figs. 4(b) and 5(b), and $\alpha = 0.5$ and $\beta = 1.0$ in Figs. 4(d) and 5(d)). Note that in the crossing regions, CFARI, represented by $\alpha = 0$, cannot find the correct crossing directions. In these examples, the errors of the proposed method can be smaller



(a) $\sigma = 0$, $\theta = 0°$

(b) $\sigma = 0$, $\theta = 15°$

(c) $\sigma = 0.06$, $\theta = 0°$

(d) $\sigma = 0.06$, $\theta = 15°$

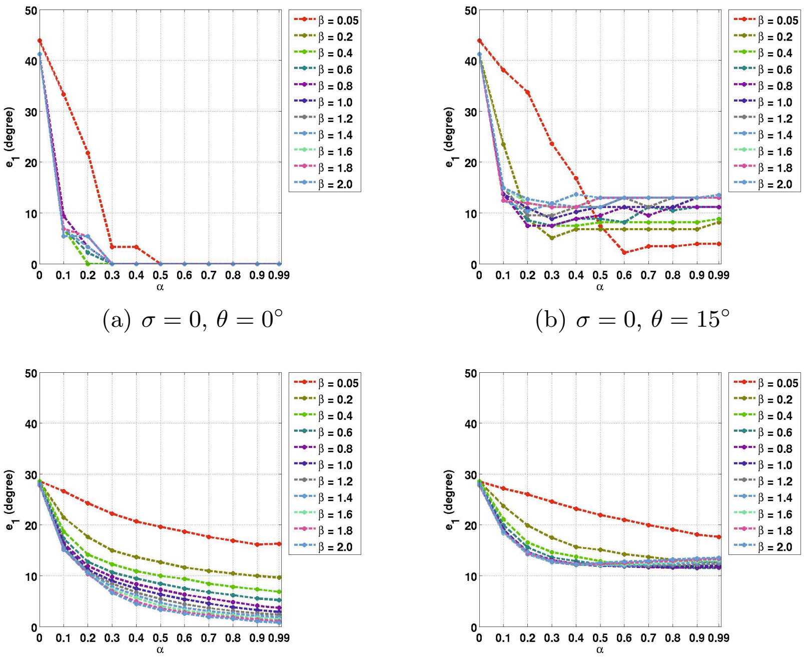
Fig. 3. Average e_2 errors in noncrossing regions with different noise level σ, PD inaccuracy θ, and the parameters of α and β

than the errors introduced in the PDs, which indicates that the proposed result is a better estimate than simply using the prior directions as the estimate.

3.2 *In Vivo* Tongue Diffusion Data

Next, we applied our method to *in vivo* tongue diffusion data of a control subject. Diffusion weighted images were acquired on a 3T MR scanner (Magnetom Trio, Siemens Medical Solutions, Erlangen, Germany) in about two minutes and 30 seconds. Each scan has 12 gradient directions and one $b0$ image. The b-value is 500 s/mm^2. The field of view (FOV) is 240 mm × 240 mm × 84 mm. The resolution is 3 mm isotropic.

To obtain PDs, we built a template by manually identifying regions of interest (ROIs) for the genioglossus (GG), the transverse muscle (T), and the vertical muscle (V) on a high resolution structural image (0.8 mm isotropic) of a subject according to [10]. T interdigitates with GG near the mid-sagittal planes and with V on lateral parts of the tongue. The $b0$ image was also acquired for this template subject in the space of the structural image. The ROIs were then subsampled to have the same resolution with the $b0$ image. Using SyN registration [11] between



(a) $\sigma = 0$, $\theta = 0°$

(b) $\sigma = 0$, $\theta = 15°$

(c) $\sigma = 0.06$, $\theta = 0°$

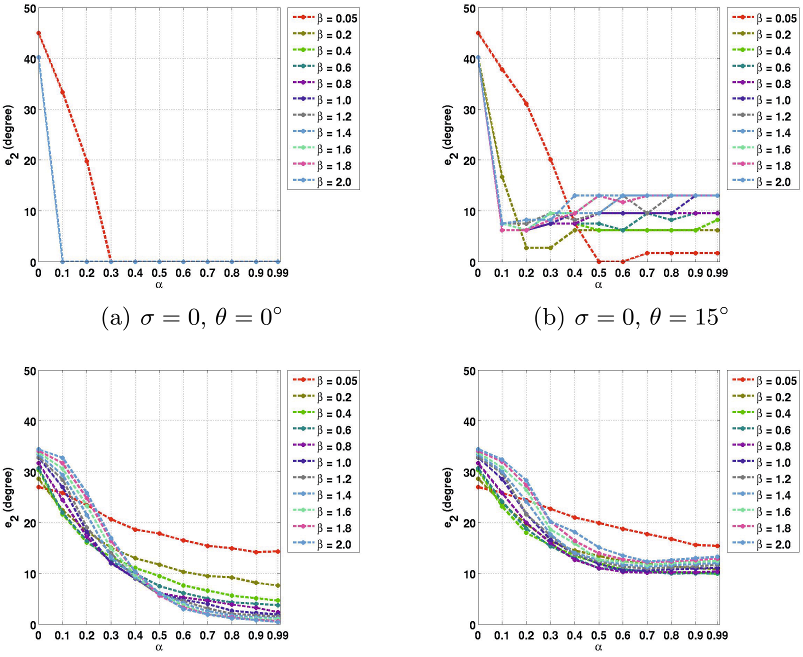
(d) $\sigma = 0.06$, $\theta = 15°$

Fig. 4. Average e_1 errors in crossing regions with different noise level σ, PD inaccuracy θ, and the parameters of α and β

$b0$ images with mutual information as the similarity metric, the template was deformed to the test subject.

Based on the deformed ROIs of GG, T, and V, PDs can be determined. GG and V are known to be fan-shaped; therefore, to calculate the PDs at each voxel (x_i, y_i, z_i) belonging to GG or V, we manually identified the origin point (x_0, y_0, z_0) of GG in the mid-sagittal slice only. Then the PD for GG or V is $\boldsymbol{w}_{\mathrm{GG/V}} = (0, y_i - y_0, z_i - z_0)$. Since T propagates transversely, we use $\boldsymbol{w}_{\mathrm{T}} = (1, 0, 0)$ as the PDs for T. An example of the PDs on the test subject is shown in Fig. 6(a). Note that in the sagittal view, left-right directions are not shown.

The proposed method was then performed with the PDs. We fixed $\beta = 1$ and tested with different α's. The result is compared with CFARI in Figs. 6(b) and 6(c). We focus on the highlighted areas in Fig. 6(a). Only directions with normalized mixture fractions $\tilde{f}_{n_i} > 0.1$ are shown. In Fig. 6(b), CFARI does not generate a good fanning pattern for GG, while by tuning α our method is able to reconstruct the fan-shaped directions. Also, in Fig. 6(c), CFARI does not produce the transverse fiber directions while in the proposed method, as α increases, transverse patterns become more obvious.



(a) $\sigma = 0$, $\theta = 0°$

(b) $\sigma = 0$, $\theta = 15°$

(c) $\sigma = 0.06$, $\theta = 0°$

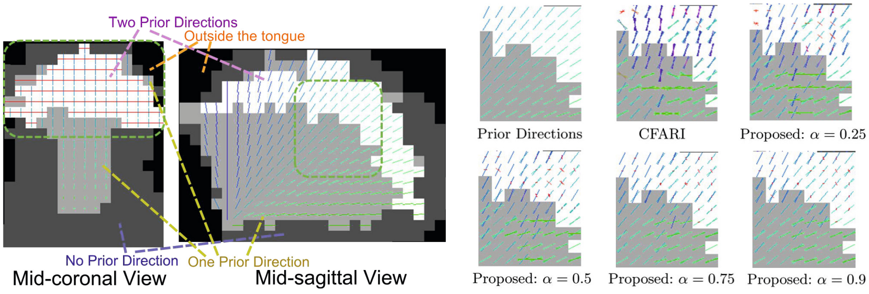
(d) $\sigma = 0.06$, $\theta = 15°$

Fig. 5. Average e_2 errors in crossing regions with different noise level σ, PD inaccuracy θ, and the parameters of α and β

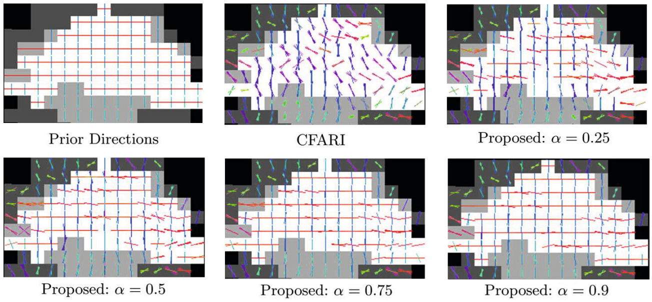
We then applied fiber tracking with the CFARI results and the proposed results (for $\alpha = 0.5$ and $\beta = 1$). We implemented a variation of INFACT tracking from [2]. The difference is that in seeding a voxel, all the reconstructed directions are used instead of only the one with the largest mixture. Seeding ROIs are placed in parts of T and GG near the mid-sagittal plane, and the results are shown in Fig. 7. Each fiber segment is color-coded by the local orientation using the standard DTI color scheme. Compared to CFARI, the proposed method reconstructs many more transverse fibers and produces a smoother fan-shaped GG.

4 Discussion

To recover crossing fiber directions, more advanced diffusion imaging, such as high angular resolution diffusion imaging (HARDI) [12] and diffusion spectrum imaging [13] (DSI), have been developed. Since HARDI and DSI usually require long acquisition time, which limits their use in clinical research, efforts have also been made to accelerate the imaging process [14]. For example, [14] reduces the scan time from 50 minutes to 17 minutes. However, in the application of tongue diffusion imaging, even accelerated imaging currently may not satisfy the scan time of around 2.5 minutes.



(a) PDs on the Test Subject (b) Fiber Directions in GG



(c) Fiber Directions in the crossing of T and GG/V

Fig. 6. Fiber directions. Results are compared between the proposed method and CFARI in (b) and (c) in the highlighted regions in (a).

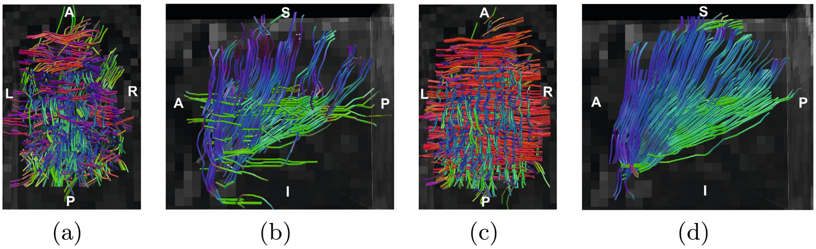


Fig. 7. Fiber tracking results: CFARI results seeded in (a) T and (b) GG; proposed results seeded in (c) T and (d) GG. T is viewed from above and GG is viewed from the left.

Like [5] and [2], we do not explicitly enforce the constraint of $||\boldsymbol{f}||_1 = 1$. As discussed in [15], the general sparse reconstruction problem (without prior knowledge) should be formulated as

$$\hat{\boldsymbol{f}} = \underset{\boldsymbol{f}\geq\boldsymbol{0},||\boldsymbol{f}||_1=1}{\arg\min} \; ||\mathbf{G}\boldsymbol{f} - \boldsymbol{y}||_2^2 + \beta||\boldsymbol{f}||_0. \tag{14}$$

The CFARI algorithm [2] can be viewed as first relaxing the constraint $||\boldsymbol{f}||_1 = 1$, then approximating the ℓ_0-norm with the ℓ_1-norm, and finally reprojecting $\boldsymbol{f}$ onto the plane $||\boldsymbol{f}||_1 = 1$. As shown in [2], the approximation is able to resolve crossing fibers. The proposed work further generalizes the approximation with weighted ℓ_1-norm using a Bayesian framework, where prior directional information is incorporated. As demonstrated in the results, the generalization can better distinguish interdigitated tongue muscles with limited gradient directions.

We have assumed a Rician noise model and approximated it with a Gaussian model. It should be noted that in the case of parallel imaging, the noise can follow a noncentral χ distribution [16]. However, in our application, the Gaussian model provides a reasonable approximation in practice.

The PDs are calculated based on the deformed muscle ROIs. An alternative way of calculation is deforming the PDs drawn on the template to the target with the deformation field. As well as the spatial position, the orientation of the PDs should also be rotated according to the deformation field, as suggested in [17]. However, we discovered that although deformable registration can provide a general location of the tracts, due to the low contrast of $b0$ images, the detailed local deformation is not necessarily accurate, leading to distorted PDs. Therefore, we choose to calculate the PDs as proposed.

The proposed method relies on the ability of specifying PDs. Because of the well organized structures of the tongue muscles, the PDs are achievable for normal subjects. When applied to patients with glossectomy, the current prior knowledge in the lesion may be misleading. Thus, a criterion for using the PDs should be decided or the PDs for patients can be determined in a different way.

Currently, the choice of α and β is empirically fixed for all the voxels. However, the weight of sparsity and prior knowledge can depend on the signal-to-noise ratio (SNR). An improvement could be to determine adaptive α and β based on the estimation of SNR. For example, the SNR can be roughly estimated using image intensities of background and foreground voxels.

5 Conclusion

We have introduced a Bayesian formulation to introduce prior knowledge into a multi-tensor estimation framework. It is particularly suited for situations where acquisitions must be fast such as in *in vivo* tongue imaging. We use a MAP framework, where prior directional knowledge and sparsity are incorporated in the prior distribution and data fidelity is ensured in the likelihood term. The problem is solved as a noise-aware version of a weighted ℓ_1-norm minimization. Experiments on a digital phantom and *in vivo* tongue diffusion data demonstrate that the proposed method can reconstruct crossing directions with limited diffusion weighted imaging.

References

1. Basser, P.J., Mattiello, J., LeBihan, D.: MR diffusion tensor spectroscopy and imaging. Biophysical Journal 66(1), 259–267 (1994)
2. Landman, B.A., Bogovic, J.A., Wan, H., ElShahaby, F.E.Z., Bazin, P.L., Prince, J.L.: Resolution of crossing fibers with constrained compressed sensing using diffusion tensor MRI. NeuroImage 59(3), 2175–2186 (2012)

3. Behrens, T.E.J., Berg, H.J., Jbabdi, S., Rushworth, M.F.S., Woolrich, M.W.: Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? NeuroImage 34(1), 144–155 (2007)
4. Peled, S., Friman, O., Jolesz, F., Westin, C.F.: Geometrically constrained two-tensor model for crossing tracts in DWI. Magnetic Resonance Imaging 24(9), 1263–1270 (2006)
5. Ramirez-Manzanares, A., Rivera, M., Vemuri, B.C., Carney, P., Mareci, T.: Diffusion basis functions decomposition for estimating white matter intravoxel fiber geometry. IEEE Transactions on Medical Imaging 26(8), 1091–1102 (2007)
6. Candes, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted ℓ_1 minimization. Journal of Fourier Analysis and Applications 14(5-6), 877–905 (2008)
7. Stejskal, E.O., Tanner, J.E.: Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. The Journal of Chemical Physics 42(1), 288 (1965)
8. Gudbjartsson, H., Patz, S.: The Rician distribution of noisy MRI data. Magnetic Resonance in Medicine 34(6), 910–914 (1995)
9. Kim, S.-J., Koh, K., Lustig, M., Boyd, S.: An efficient method for compressed sensing. In: IEEE International Conference on Image Processing, ICIP 2007, vol. 3, pp. 111–117. IEEE (2007)
10. Takemoto, H.: Morphological analyses of the human tongue musculature for three-dimensional modeling. Journal of Speech, Language, and Hearing Research 44(1), 95–107 (2001)
11. Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis 12(1), 26–41 (2008)
12. Tuch, D.S., Reese, T.G., Wiegell, M.R., Makris, N., Belliveau, J.W., Wedeen, V.J.: High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. Magnetic Resonance in Medicine 48(4), 577–582 (2002)
13. Wedeen, V.J., Hagmann, P., Tseng, W.Y.I., Reese, T.G., Weisskoff, R.M.: Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging. Magnetic Resonance in Medicine 54(6), 1377–1386 (2005)
14. Bilgic, B., Setsompop, K., Cohen-Adad, J., Yendiki, A., Wald, L.L., Adalsteinsson, E.: Accelerated diffusion spectrum imaging with compressed sensing using adaptive dictionaries. Magnetic Resonance in Medicine 68(6), 1747–1754 (2012)
15. Daducci, A., Van De Ville, D., Thiran, J.P., Wiaux, Y.: Reweighted sparse deconvolution for high angular resolution diffusion MRI. arXiv preprint arXiv:1208.2247 (2012)
16. Dietrich, O., Raya, J.G., Reeder, S.B., Ingrisch, M., Reiser, M.F., Schoenberg, S.O.: Influence of multichannel combination, parallel imaging and other reconstruction techniques on MRI noise characteristics. Magnetic Resonance Imaging 26(6), 754–762 (2008)
17. Alexander, D.C., Pierpaoli, C., Basser, P.J., Gee, J.C.: Spatial transformations of diffusion tensor magnetic resonance images. IEEE Transactions on Medical Imaging 20(11), 1131–1139 (2001)

Optimal Joint Segmentation and Tracking of *Escherichia Coli* in the Mother Machine

Florian Jug[1], Tobias Pietzsch[1], Dagmar Kainmüller[1], Jan Funke[2], Matthias Kaiser[3], Erik van Nimwegen[3], Carsten Rother[4], and Gene Myers[1]

[1] Max Planck Institute of Molecular Cell Biology and Genetics, Germany
[2] Institute of Neuroinformatics, Univerity Zurich / ETH Zurich, Switzerland
[3] Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Switzerland
[4] Computer Vision Lab Dresden, Technical University Dresden, Germany

Abstract. We introduce a graphical model for the joint segmentation and tracking of *E. coli* cells from time lapse videos. In our setup cells are grown in narrow columns (growth channels) in a so-called "Mother Machine" [1]. In these growth channels, cells are vertically aligned, grow and divide over time, and eventually leave the channel at the top. The model is built on a large set of cell segmentation hypotheses for each video frame that we extract from data using a novel parametric max-flow variation. Possible tracking assignments between segments across time, including cell identity mapping, cell division, and cell exit events are enumerated. Each such assignment is represented as a binary decision variable with unary costs based on image and object features of the involved segments. We find a cost-minimal and consistent solution by solving an integer linear program. We introduce a new and important type of constraint that ensures that cells exit the Mother Machine in the correct order. Our method finds a globally optimal tracking solution with an accuracy of $> 95\%$ (1.22 times the inter-observer error) and is on average $2 - 11$ times faster than the microscope produces the raw data.

1 Introduction

The *Mother Machine* [1] is a microfluidic device designed to study live bacteria. It allows the observation of growth and division of the progeny of single "mother" cells over many generations using time lapse microscopy. Figure 1 illustrates the Mother Machine and the respective image data. In such data, individual bacteria need to be *tracked* over time. Tracking consist of two equally important tasks: (*i*) cells need to be segmented in each frame, and (*ii*) all segments of the same cell need to be linked between frames. Tracking large numbers of cells under different environmental conditions will allow biologists to better understand the stochastic dynamics of gene expression within living cells. Respective high throughput studies of cells in the Mother Machine would be greatly facilitated if the tracking task would be automated.

Many existing automated tracking systems perform the tasks of segmentation and linkage one after another to reduce overall model complexity and runtime [2].

M.J. Cardoso et al. (Eds.): BAMBI 2014, LNCS 8677, pp. 25–36, 2014.

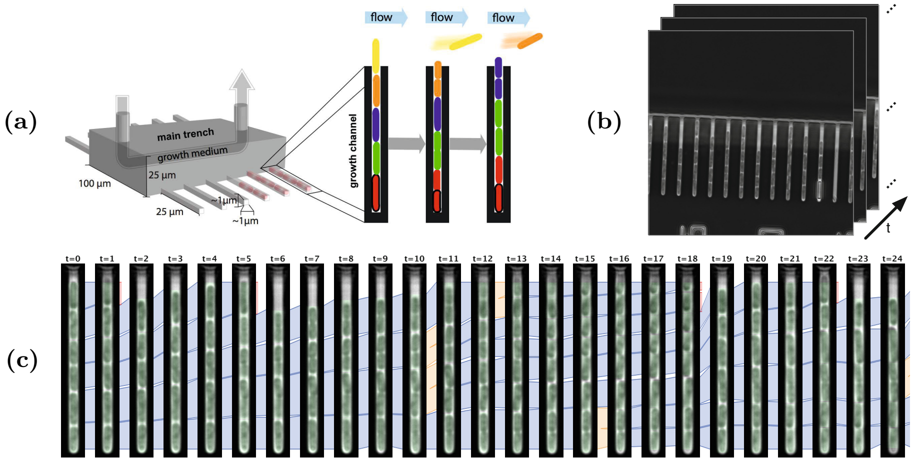


Fig. 1. (a) Illustration of the "Mother Machine", a microfluidic device built to understand dynamic processes in *E. coli*. Individual 'growth channels' (narrow tubes, just wide enough for hosting a row of bacteria) are imaged every minute. **(b)** Raw images. **(c)** One growth channel in the first 25 frames of a time-lapse movie. A tracking is shown between frames, with mapping assignments in blue, division assignments in yellow, and exit assignments in red.

Model complexity is typically reduced even further by performing linkage in a locally optimal, *greedy* fashion [2], i.e. frame by frame, never considering the whole time series at once.

However, *globally optimal joint segmentation and linkage* can be achieved by so-called *Assignment Models* [3,4,5,6,7]. Assignment models pose the linkage problem as a global energy minimization task, where the energy is that of a graphical model (*factor graph*). Binary variables represent possible links (called *assignments*), with respective unary potentials capturing their plausibility. Higher order factors encode *continuity constraints*, that describe which link sequences form structurally sound tracks. Assignment models can elegantly handle an excess of non overlapping *segment hypotheses*[1]. The only extra ingredient are additional unary factors assigning costs to all segment hypotheses. Energy minimization in such a model yields globally optimal, joint segmentation and tracking. The respective optimization task can be solved with existing discrete optimization methods [5,6,7].

A good assignment model should allow as many different segmentation hypotheses as possible to avoid missing segments (i.e. good *recall*). To this end, Kausler et al. [5] and Schiegg et al. [7] allow for an *over-segmentation* per time-frame. To be yet more robust against missed segments (false-negatives), Schiegg et al. propose a

[1] Superfluous segments do not have to be linked between frames but can be filtered by the tracking engine.

method capable of dealing with occasional under-segmentations. Funke et al. [6] introduced a model capable of dealing with a large pool of partially *conflicting* (overlapping) segment hypotheses per frame. Their model filters a conflict-free subset by introducing adequate higher order factors, called *tree constraints.* The work we presented here follows this "hypotheses-rich" approach of Funke et al. [6].

But in order to be as specific as possible for a given task (i.e. good *precision*), assignment models should be designed to restrict the space of possible solutions as much as possible. So far, relatively *generic* prior knowledge on cell movement and proliferation has been encoded into assignment models: Cells can be kept from moving too far between time frames; They can be allowed to divide but not merge; They can be kept from dividing into more than two, and kept from appearing from nowhere. However, none of the previously published assignment models captures a particular kind of prior knowledge that is important for cells in the Mother Machine, namely the *total order* of cells within growth channels which has to be maintained at any time.

The main technical contribution of this paper is a novel type of higher order factors which are concerned with the order of cells within growth channels. We call these factors *exit constraints.* We show that exit constraints considerably improve tracking accuracy in the Mother Machine (see Section 4.2). Another contribution is a new approach for generating nested segmentation hypotheses which outperforms previous approaches. The idea is to combine the benefits of parametric max flow [8] and random forest classifiers [9]. The random forest is used to improve the separation of recently divided cells which are otherwise hard to tell apart (see Section 3).

Our proposed assignment model can solve the problem of tracking cells in the Mother Machine with an error rate of 4.8%, which is only 1.22 times the inter-observer error (see Section 5). Hence our system renders high throughput imaging and tracking of bacteria in the Mother Machine possible.

2 Microscopic Setup and Data Preprocessing

The Mother Machine consists of a main trench and dead end growth channels that host the bacterial cells (see Figure 1). The width of the growth channels is chosen such that each of them fits only a single bacterial cell, thereby forcing the growing cells into a linear array. A constant flow in the main trench leads to continuous diffusion of nutrients and removes cells that emerge from the growth channels. Experiments are imaged by an inverted microscope equipped with an incubator. Images are taken every minute using a 100x objective.

Raw data from the microscope undergoes a few simple preprocessing steps. Two of those are of particular importance. First, movie frames are rotated into an upright orientation, because growth channels are usually tilted by up to ±45°, see Figure 1(b). To determine the tilt angle, we smooth each image row, collect local maxima, and fit straight lines through each growth channel.

In a second step we correct for uneven background, caused by uneven lighting and different material thicknesses of the Mother Machine itself. For each growth

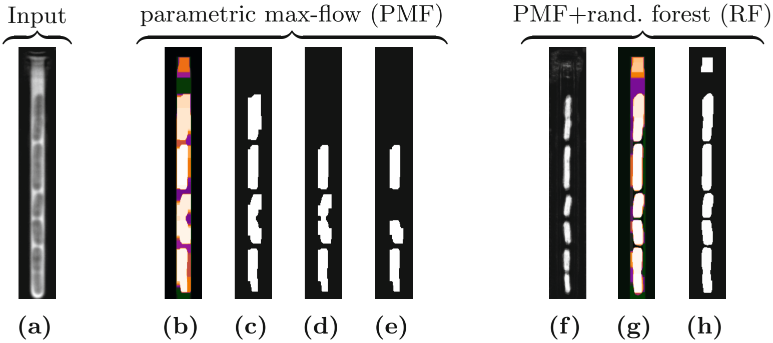


Fig. 2. Parametric max-flow based generation of segmentation hypotheses with and without using a random forest classifier (RF) to modulate unary and binary potentials. **(a)** Image to be segmented. **(b-e)** Results obtained using parametric max-flow. **(f-h)** Results when potentials are modified by a trained RF. **(b,g)** all graph-cut segmentations given by parametric max-flow. The color of a pixel is determined by the number of times this pixel is classified as foreground. **(c,d,e)** three graph cut solutions (of 5176). **(f)** probability map given by RF, trained to over-emphasize gaps between cells. **(h)** single graph-cut containing the correct segmentation and a false positive at the very top.

line we evaluate the background intensity at each height by averaging the intensities of automatically selected local image patches from within the "empty" areas to either side. This intensity is subtracted from each growth-channel pixel at the given height.

Images for each indivual growth-channel are then cropped from the preprocessed image; an example is shown in Figure 2(a).

3 Segmentation Methods

Automated tracking approaches face the challenge that each segmentation error directly translates to at least one tracking error. Assignment models tackle this problem by not committing to a segment for as long as possible. Instead, an excess of potentially conflicting (overlapping) segment hypotheses is created and the model filters the best consistent subset [6]. Below we introduce 3 segmentation methods we use for this purpose.

3.1 Thresholding and Component Trees (CT)

The first segmentation methods we use is an intensity thresholding technique similar to [10,11]. Any threshold yields a binary image from which connected foreground components can be extracted. When the threshold level is gradually raised foreground components grow until they eventually merge. This allows for grouping all components for all thresholds in a tree data structure, called a *component tree*. Nodes in the component tree, i.e. individual segmented regions rather than a global segmentation, correspond to segment hypotheses.

3.2 Parametric Max-Flow (PMF)

Parametric max-flow [8] is a graph-cut formulation with an additional, additive parameter λ. This parameter linearly scales the unary costs, leading to different segmentation results. The corresponding energy can be formalized as

$$E^{\lambda}(\mathbf{x}) = \sum_{u \in V} (a_u + \lambda) x_u + \sum_{(u,v) \in E} f_{uv}(x_u, x_v), \tag{1}$$

where $\mathbf{x}$ is a vector of binary variables $x_u \in \{0, 1\}$, f_{uv}'s are hand tuned and submodular, $\lambda \in I \subseteq \mathbb{R}$, and $G = (V, E)$ is an undirected graph, in our case the 4-connected grid graph on the pixels of each frame. Values $x_u = 0$ and $x_u = 1$ represent pixel labels "foreground" (cell) and "background", respectively. Unary costs a_u for a pixel u depend on measured intensity distributions for foreground and background pixels. Pairwise costs $f_{uv}(x_u, x_v)$ are inversely proportional to the intensity gradient between pixels u and v. More details can e.g. be found in [12]. The work by Kolmogorov et al. [8] offers an efficient way to compute all solutions for $E^{\lambda}(\mathbf{x})$ for all $\lambda \in \mathbb{R}$, which is a finite and nested set, typically counting between 10 and 10000 solutions.

Like components for increasing threshold values, also the components obtained by increasing λ are monotonically growing. Hence, we can again store all segment hypotheses in a tree. The benefit of PMF over thresholding alone is the additional smoothing that comes with the graph-cut formulation.

3.3 Parametric Max-Flow and Random Forest (PMFRF)

Since missing segments immediately lead to bad tracking performance we combine parametric max-flow and a trained random forest classifier (RFC). This predictor for cell vs. background pixels $P(x_u)$ is trained using the Fiji plugin "Trainable Weka-Segmentation" [13] and manually tuned to pick up even very small clefts between, for example, freshly divided cells. This is done to avoid undersegmentation in cases where the cleft between adjacent cells is not clearly visible (false positives can always be filtered by the model later on, but false negatives translate directly to tracking errors). For the data presented here we trained the RFC on only 3 raw images that where taken from a different raw dataset.

The probability map P for the 'cell'-class is used to modify the costs a_u and $f_{uv}(x_u, x_v)$ of Equation (1) as follows (see Figure 2 for an illustration):

$$a_u^{\text{trained}} = a_u \cdot P(x_u), \text{ and} \tag{2}$$

$$f_{uv}^{\text{trained}}(x_u, x_v) = f_{uv}(x_u, x_v) \cdot (1 - |P(x_u) - P(x_v)|) \, . \tag{3}$$

4 A Graphical Model for Segmentation and Tracking

We choose the language of factor graphs to describe a model for joint segmentation and tracking in Mother Machine datasets. Here, segmentation consists of selecting a consistent subset of the segment hypotheses $H^{(t)}$ for each time-point. See Figure 3 for an illustration. We use a *factor graph* $\mathcal{FG} = (\mathcal{V}, \mathcal{F}, \mathcal{E})$ with $\mathcal{V}$ being a set of binary variables or *variable nodes*, $\mathcal{F}$ being a set of factors or *factor nodes*, and $\mathcal{E} \subset \mathcal{V} \times \mathcal{F}$ [14].

Variable Nodes. The variable nodes $\mathcal{V} = \mathcal{H} \cup \mathcal{A}$ comprise *segmentation variables* $\mathcal{H} = \bigcup_{t=1}^{T} H^{(t)}$ and *assignment variables* $\mathcal{A} = \bigcup_{t=1}^{T-1} A^{(t)}$.

Each binary segmentation variable $h^{(t)} \in H^{(t)}$ indicates whether a particular segment hypothesis at time-point t is choosen as part of the solution. Assignment variables $a^{(t)} \in A^{(t)}$ link segment hypotheses at time-point t to segment hypotheses at time-point $t+1$. We distinguish three types of assignment variables.

Mapping assignments: A mapping assignment $a^{(t)}_{i \mapsto j}$ connects two segment hypotheses $h^{(t)}_i$ and $h^{(t+1)}_j$. It indicates that these segments correspond to the same segmented cell that is tracked between time-points t and $t+1$.

Division assignments: A division assignment $a^{(t)}_{i \div jk}$ connects segment hypothesis $h^{(t)}_i$ to $h^{(t+1)}_j$ and $h^{(t+1)}_k$. It indicates that these segments correspond to a cell division event, where one segmented cell at time-point t divides into two daughter cells at $t+1$.

Exit assignments: An exit assignment $a^{(t)}_{\perp i}$ is only connected to one segment hypothesis $h^{(t)}_i$. It indicates that this segment corresponds to a segmented cell at time-point t that is spilled out on top of a growth line at time-point $t+1$.

Factor Nodes. Factor nodes connect to one or more variable nodes, assigning a potential to each joint configuration of these variables. The factor nodes $\mathcal{F}$ comprise unary factors and higher order factors. Unary factors $f(v)$ are connected to each binary variable $v \in \mathcal{V}$, capturing the plausibility that v is active given the data. Formally we define

$$-\ln f(v) = \begin{cases} 0 & \text{if } v = 0 \\ c_v & \text{if } v = 1, \end{cases} \tag{4}$$

c_v is the cost for including the respective segmentation or assignment variable in the solution. These costs are derived from (image) features as described in the next subsection. Structural constraints are expressed as n-ary factors for which $-\ln f(var(E)) = 0$ if E holds, and ∞ otherwise. where E are (in-)equalities on the set of variables $var(E)$ connected to the factor. The constraints formalized by these (in-)equalities prohibit solutions involving conflicting segmentation hypotheses or assignments that are inconsistent with the selected segmentation.[2] Constraints are described in Section 4.2.

[2] Such inconsistent solutions correspond to events with infinite costs or 0 probability.

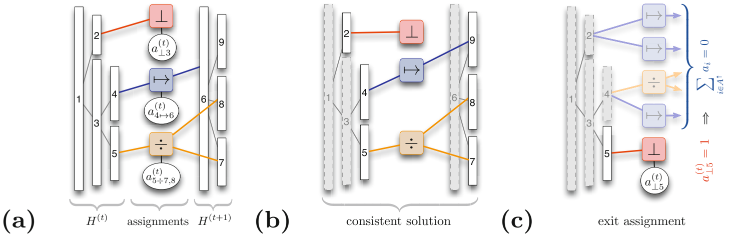


Fig. 3. Overview of the proposed model. Possibly contradictory segmentation hypotheses for two time frames are shown in gray. Between frames, binary assignments variables of three types are enumerated: continuations (blue), divisions (orange), and exits (red). Selecting an assignment variable jointly selects the involved segmentation hypotheses and assigns them to each other.

4.1 Costs

All costs c_v corresponding to activating a variables v are defined according to the following considerations.

We define negative costs for segmentation variables in order to provide an incentive to activate segment hypotheses. Otherwise the trivial solution of ‘seeing’ only empty growth lines, corresponding to a total cost of 0, would be optimal.

We derive segmentation costs from the image intensities along the pixel row at the center of the growth line with the following intuition in mind. A strong gradient on the upper and lower border of a hypothesis increases the likelihood of it being a correct segment and therefore lowers the cost. A strong gradient in the interior of a hypothesis decreases the likelihood (increases the cost) because it suggests that it might contain several cells. Finally, we scale the cost by the size of the segment hypothesis. The rationale for this is that we want to favor hypotheses that explain a larger part of the image in cases where equal support is given by the previously mentioned gradient based measures.

The costs for assignment variables are derived from the positions and sizes of segment hypotheses connected by this assignment. As time progresses from one frame in a given time-lapse movie to the next, we expect an average change in the size and position of a cell.

For mapping assignments we compare the segment sizes and centroids at time points t and $t+1$. The cost for a mapping assignment is given by a suitably defined function that reflects how unlikely certain deviations from the expected size change and the expected centroid shift really are. This is actually a very natural way of utilizing the knowledge of biological experts.

Costs for division assignments are defined similarly. Here, a segment at timepoint t is linked to two (adjacent) segments at $t+1$. In addition we know that a dividing cell usually distributes its volume equally to its daughters. We compute size and centroid from the union of the two segment hypotheses at $t+1$ and compute the cost as described for mapping assignments, plus some additional cost for unequally sized segments at $t+1$.

Last but not least we have to define costs for exit assignments. With the rationale in mind that an early exit assignment already leads to not segmenting this cell in future time-points (thereby not 'earning' the corresponding negative cost) we assign 0 cost to all exit assignments.

4.2 Constraints

Tree Constraints. It is important to note that sequential thresholding as well as parametric max-flow respectively yield a monotonic sequence of solutions, inducing a partial order on the segment hypotheses to form a tree $(H^{(t)}, \supset)$.

We say that segment hypotheses $h_i^{(t)} \supset h_j^{(t)}$ are conflicting because they offer mutually exclusive interpretations of (parts of) the image data. Of all segment hypotheses on a branch $h_1^{(t)} \supset \cdots \supset h_n^{(t)}$, only one can be simultaneously valid because we seek an assignment of each image pixel to exactly one segment (or background). Tree constraints enforce that conflicting segment variables cannot be simultaneously active. This is formalized in the set of inequalities

$$\forall t \in \{1, \ldots, T\},\ \forall \pi \in \mathcal{P}(H^{(t)}) : \sum_{h^{(t)} \in \pi} h^{(t)} \leq 1 \tag{5}$$

where $\mathcal{P}(H^{(t)})$ is the set of all paths π from the root node in $(H^{(t)}, \supset)$ to any of its leaf nodes.

Continuity Constraints. Continuity constraints enforce consistency between segmentation and assignment variables. If a segment hypothesis is selected, exactly one of the assignments entering it from the previous time-point, and exactly one of the assignments leaving it towards the next time-point must be selected as well. If a segment hypothesis is not selected, neither must any of these assignments be selected. This is formalized as the following sets of constraints. For the entering assignments we have

$$\forall t \in \{2, \ldots, T\},\ \forall h^{(t)} \in H^{(t)} : \sum_{a^{(t-1)} \in \mathit{\Gamma}_{\mathrm{L}}\left(h^{(t)}\right)} a^{(t-1)} = h^{(t)} \tag{6}$$

where the *left neighborhood* $\mathit{\Gamma}_{\mathrm{L}}(h)$ is the set of all assignments entering h from the previous time-point. That is, $\mathit{\Gamma}_{\mathrm{L}}\left(h_i^{(t)}\right)$ contains assignments $a_{j \mapsto i}^{(t-1)}$, $a_{j \div ik}^{(t-1)}$, and $a_{j \div ki}^{(t-1)}$ (for all j, k). Similarly, for the assignments leaving to the next time-point we have

$$\forall t \in \{1, \ldots, T-1\},\ \forall h^{(t)} \in H^{(t)} : \sum_{a^{(t)} \in \mathit{\Gamma}_{\mathrm{R}}\left(h^{(t)}\right)} a^{(t)} = h^{(t)} \tag{7}$$

where the *right neighborhood* $\mathit{\Gamma}_{\mathrm{R}}\left(h\right)$ is the set of all assignments leaving h. That is, $\mathit{\Gamma}_{\mathrm{R}}\left(h_i^{(t)}\right)$ contains assignments $a_{\perp i}^{(t)}$, $a_{i\mapsto j}^{(t)}$, and $a_{i\div jk}^{(t)}$ (for all j,k).

Exit Constraints. One of the main contributions of this article is the introduction of this specific type of constraint. It is obvious that cells can only exit the growth line at the very top. A cell in the middle of a growth line can impossibly be spilled out without all other cells above it being spilled out as well. Let us denote by $A_{\uparrow}(h^{(t)}) \subset A^{(t)}$ the set of mapping and division (but not exit) assignments that are leaving hypotheses located strictly above $h^{(t)}$. If the exit assignment is chosen for segment h, then none of the assignments in $A_{\uparrow}(h)$ can be active. (See Figure 3(c) for an illustration.) However, if the exit assignment for h is not chosen, any number of these assignments might be active. We express this as the set of inequalities

$$\forall t \in \{1,\ldots,T-1\},\ \forall h_i^{(t)} \in H^{(t)} : |H^{(t)}| \cdot a_{\perp i}^{(t)} + \sum_{a \in A_{\uparrow}(h_i^{(t)})} a \leq |H^{(t)}|. \tag{8}$$

Note that, in combination with the continuity constraints (7), this forces all active segments above an exiting hypothesis to exit as well, thereby maintaining the linear order of cells in the mother machine also in our tracking results.

To quantify the importance of exit constraints we removed all exit contraints from our model and tracked all available datasets. We then compared the results to ground truth as explained in Section 5. Error rates increased to 225% (on average to 123%), clearly hinting at the importance of these constraints.

4.3 Eliminating Segmentation Variables

Considering the costs and constraints defined above it can be seen that segmentation variables are redundant in the formulation of the factor graph. The continuity equality (7) provides a definition for each segmentation variable in terms of a sum over a set of assignment variables. Plugging these definitions into (5), and replacing (6) and (7) by

$$\forall t \in \{2,\ldots,T-1\},\ \forall h^{(t)} \in H^{(t)} : \sum_{a^{(t-1)} \in \mathit{\Gamma}_{\mathrm{L}}\left(h^{(t)}\right)} a^{(t-1)} - \sum_{a^{(t)} \in \mathit{\Gamma}_{\mathrm{R}}\left(h^{(t)}\right)} a^{(t)} = 0 \tag{9}$$

we can eliminate segmentation variables from the constraints.[3]

Similarly, the costs c_h can be dropped, and added to the cost of each exiting assigment c_a, where the constraints guarantee that at most one of these is active.[4]

[3] One might fear that by replacing h by a sum over assignment variables might loose the restriction that h is binary Note, however, that this is now effectively ensured by the tree constraints (5) (with $h^{(t)}$ replaced).

[4] The costs of segmentation hypotheses $h^{(T)}$, which have no exiting assignments, are added to each entering assignment instead.

4.4 Finding The Globally Optimal Solution

A globally optimal segmentation and tracking is provided by a MAP (maximum *a posteriori* probability) or, equivalently, minimum energy solution of the factor graph. This amounts to finding a conflict-free variable assignment (not violating any constraint) with minimal summed cost.

Similarly to [5,6,7] we formulate the problem as an integer linear program (ILP) [15]: The cost of a conflict-free solution yields the linear objective we wish to minimize[5]. The feasible space is restricted to conflict-free solutions by the linear constraints discussed in Section 4.2 (and additional constraints $0 \leq v \leq 1$ to ensure that all variables $v \in \mathbb{Z}$ are binary). This approach guarantees to find a globally optimal solution, the worst-case complexity is though exponential. In all our experiments we observe runtimes (for ILP solving alone) in the range is a couple of seconds only. See also Figure 5.

We use the off-the-shelf ILP solver Gurobi™ to find the optimal solution.

5 Results

We tested our model on 2 movies containing a total of 21 datasets (growth channels). In order to measure the error of our fully automated tracking pipeline we have manually created ground truth (GT) for all given datasets.

We count (*i*) *segmentation mismatch*, and (*ii*) *tracking errors*. For both we greedily match all segments in a given solution with the corresponding segments in the GT. Segmentation mismatch is measured by adding offsets between uppermost pixels and lowermost pixels in each matched segment pair.

The tracking error counts over- and undersegmentations, computed by comparing the number of active segments at any given time-point in solution and the GT, and assignment-type mismatches. For those we count type-mismatches for all right-assignments (assignments towards next time-point) associated to pairs of matched segments. Note that this is a fairly pessimistic measure where errors that would intuitively be counted as one mistake are counted multiple times[6].

Figure 4 shows the results of the ground truth comparison. The first three columns in each box-plot show how the fully automated solutions compare to GT. Each column corresponds to one of the segmentation methods introduced in Section 3. The last column shows an inter-observer reliability measure.

The inter-observer reliability tells us about how much homogeneity, or consensus, there is to expect when different users create "ground truth" for the same data. We gave the automatically generated PMFRF solution and a interactive tool to 2 users, asking them to to fix all errors. We then compared their results to GT in the same way we described above. See Figure 5 for a detailed comparison of runtimes for the fully automated pipeline.

[5] It is easily seen that the summed cost is a linear function by writing it as the inner product of the vectors of all binary variables and costs, $\langle (v_1, \ldots, v_n), (c_{v_1}, \ldots, c_{v_n}) \rangle$.

[6] An early exit assignment is once counted as assignment-type mismatch and in all future time-points still containing this cell as undersegmentations.

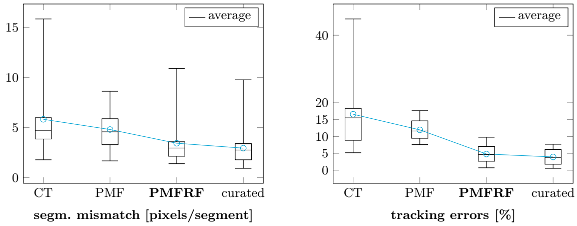


Fig. 4. Error measures for all 21 datasets. (Abbr.: CT→'component tree'; PMF→'parametrix max-flow'; PMFRF→PMF+trained random forest.) Left panel shows how well the chosen segments match to ground truth. We compare the pixel distance between the uppermost and lowermost segmented pixels between each segments and its corresponding ground truth segment. The right panel shows the fraction of assignments that do not match to ground truth.

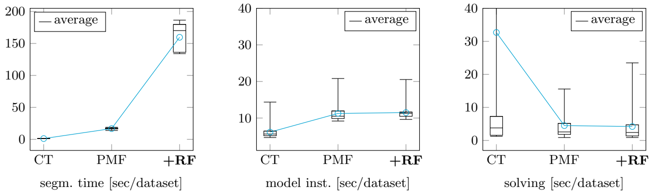


Fig. 5. Runtime for segmentation, model instantiation, and model solving. Shown times are in 'wall-time' seconds per dataset. We used a quadcore MacBook Pro Retina (Fall 2012). An excessive filter bank is main reason for slow RFs.

6 Summary and Discussion

We showed how cell tracking in the Mother Machine can be addressed using an adequately formulated assignment model. In order to achieve low error rates we needed to extend existing models [5,6,7] by additional constraints concerned with the linear order of cells in the Mother Machine and a specialized method to create nested segment hypotheses using a parametric max-flow formulation and trained random forests classifiers. Automated tracking and segmentation quality reaches a level that lies within a factor of 1.1 compared to the inter-observer variability we measured. Our system will be freely available open source software, enabling groups around the world to analyze cell cultured in the Mother Machine.

With this paper we contribute to a recent trend of formulating tracking problems as global optimization problems in the spirit of graphical models. We predict that the capabilities of assignment models is by far not reached yet.

Future extensions will focus on several important aspects such as (*i*) further increasing the set of segment hypotheses, thereby generalizing the concept of conflict trees to more general conflict graphs, (*ii*) development of more generic and task specific higher order factors that will capture ever more expert domain knowledge and therefore lead to better automated results, (*iii*) parametrization and parameter training of used cost functions, for example by means of structured learning, and (*iv*) alternative solving strategies, either by means of divide-and-conquer like dual decomposition schemes or, means of approximate inference methods, or suitable combinations.

The last mentioned point will become increasingly important with growing problem instances and the need for interactive proofreading and data curation interfaces.

Acknowledgments. This work was supported by the German Federal Ministry of Research and Education (BMBF) under the funding code 031A099.

References

1. Wang, P., Robert, L., Pelletier, J., Dang, W., Taddei, F., Wright, A., Jun, S.: Robust growth of E. coli. Current Biology 20(12), 1099–1103 (2010)
2. Jug, F., Pietzsch, T., Preibisch, S., Tomancak, P.: Bioimage informatics in the context of Drosophila research. Methods (2014)
3. Padfield, D., Rittscher, J., Roysam, B.: Coupled Minimum-Cost Flow Cell Tracking. In: Prince, J.L., Pham, D.L., Myers, K.J. (eds.) IPMI 2009. LNCS, vol. 5636, pp. 374–385. Springer, Heidelberg (2009)
4. Padfield, D., Rittscher, J., Roysam, B.: Coupled minimum-cost flow cell tracking for high-throughput quantitative analysis. Medical Image Analysis 15(4), 650–668 (2011)
5. Kausler, B.X., et al.: A Discrete Chain Graph Model for 3d+t Cell Tracking with High Misdetection Robustness. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 144–157. Springer, Heidelberg (2012)
6. Funke, J., Anders, B., Hamprecht, F., Cardona, A., Cook, M.: Efficient automatic 3D-reconstruction of branching neurons from EM data. In: CVPR. IEEE (2012)
7. Schiegg, M., Hanslovsky, P., Kausler, B., Hufnagel, L.: Conservation Tracking. In: ICCV (2013)
8. Kolmogorov, V., Boykov, Y., Rother, C.: Applications of parametric maxflow in computer vision. In: ICCV, pp. 1–8. IEEE (2007)
9. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
10. Jones, R.: Component trees for image filtering and segmentation. In: IEEE Workshop on Nonlinear Signal and Image Analysis (1997)
11. Nistér, D., Stewénius, H.: Linear Time Maximally Stable Extremal Regions. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 183–196. Springer, Heidelberg (2008)
12. Blake, A., Kohli, P., Rother, C.: Markov Random Fields for Vision and Image Processing. MIT Press (2011)
13. Arganda-Carreras, I., Cardona, A., Kaynig, V., Schindelin, J.: Trainable weka segmentation (May 2011), `http://fiji.sc/Trainable_Weka_Segmentation`
14. Frey, B., Kschischang, F., Loeliger, H., Wiberg, N.: Factor graphs and algorithms. In: Proceedings of the Annual Allerton Conference on Communication Control and Computing, vol. 35, pp. 666–680 (1997)
15. Schrijver, A.: Theory of Linear and Integer Programming. J. Wiley & Sons (1998)

Physiologically Informed Bayesian Analysis of ASL fMRI Data

Aina Frau-Pascual[1,3], Thomas Vincent[1], Jennifer Sloboda[1],
Philippe Ciuciu[2,3], and Florence Forbes[1]

[1] INRIA, MISTIS, Grenoble University, LJK, Grenoble, France
[2] CEA/DSV/I^{2}BM NeuroSpin center, Bât. 145, 91191 Gif-sur-Yvette, France
[3] INRIA, Parietal, 91893 Orsay, France

Abstract. Arterial Spin Labelling (ASL) functional Magnetic Resonance Imaging (fMRI) data provides a quantitative measure of blood perfusion, that can be correlated to neuronal activation. In contrast to BOLD measure, it is a direct measure of cerebral blood flow. However, ASL data has a lower SNR and resolution so that the recovery of the perfusion response of interest suffers from the contamination by a stronger hemodynamic component in the ASL signal. In this work we consider a model of both hemodynamic and perfusion components within the ASL signal. A physiological link between these two components is analyzed and used for a more accurate estimation of the perfusion response function in particular in the usual ASL low SNR conditions.

1 Introduction

Arterial Spin Labelling (ASL) [1] provides a direct measure of cerebral blood flow (CBF), overcoming one of the most important limitations of Blood Oxygen Level Dependent (BOLD) signal [2]: BOLD contrast cannot quantify cerebral perfusion. In contrast to BOLD, ASL is able to provide a measure of baseline CBF as well as quantitative CBF signal changes in response to stimuli presented to any volunteer in the scanner during an experimental paradigm. Hence, ASL enables the comparison of CBF changes between experiments and subjects (healthy vs patients) making its application to clinics feasible. In addition, ASL signal localization is closer to neural activity. ASL has already been used in clinics in steady-state for instance for probing CBF discrepancy in pathologies like Alzheimer's disease and stroke, but its use in the functional MRI context has been limited so far. Despite ASL advantages, its main limitation lies in its low Signal-to-Noise Ratio (SNR), which, together with its low temporal and spatial resolutions, makes the analysis of such data more challenging.

According to [3,4], ASL signal has been typically analyzed with a general linear model (GLM) approach, accounting for a BOLD component mixed with the perfusion component. In such a setting both the hemodynamic response function (HRF or BRF for BOLD response function) and perfusion response function (PRF) are assumed to be the same and to fit the canonical BRF shape. In contrast, an adaptation of the Joint-Detection estimation (JDE) framework [5]

M.J. Cardoso et al. (Eds.): BAMBI 2014, LNCS 8677, pp. 37–48, 2014.

to ASL data has been proposed in [6,7] to separately estimate BRF and PRF shapes, and implicitly consider the control/label effect which, as stated in [4], increases the sensitivity of the analysis compared to differencing approaches. Although this JDE extension provides a good estimate of the BRF, the PRF estimation remains much more difficult because of the noisier nature of the perfusion component within the ASL signal. In the past decade, physiological models have been described to explain the physiological changes caused by neural activity. In [8,9], neural coupling, which maps neural activity to ensuing CBF, and the *Balloon model*, which relates CBF to BOLD signal, have been introduced. These models describe the process from neural activation to the BOLD measure, and the impact of neural activation on other physiological parameters.

Here, we propose to rely on these physiological models to derive a tractable linear link between perfusion and BOLD components within the ASL signal and to exploit this link as a prior knowledge for the accurate and reliable recovery of the PRF shape in functional ASL data analysis. This way, we refine the separate estimation of the response functions in [6,7] by taking physiological information into consideration. The structure of this paper goes as follows: the physiological model and its linearization to find the PRF/BRF link are presented in section 2. Starting then from the ASL JDE model described in section 3, we extend the estimation framework to account for the physiological link in section 4. Finally, results on artificial and real data are presented and discussed in sections 5-7.

2 A Physiologically Informed ASL/BOLD Link

Our goal is to derive an approximate physiologically informed relationship between the perfusion and hemodynamic response functions so as to improve their estimation in a JDE framework [6,7]. We show in this section that, although this relationship is an imperfect link resulting from a linearization, it provides a good approximation and allows to capture important features such as a shift in time-to-peak from one response to another. For a physiologically validated model, we use the extended balloon model described below.

2.1 The Extended Balloon Model

The Balloon model was first proposed in [10] to link neuronal and vascular processes by considering the capillary as a balloon that dilates under the effect of blood flow variations. More specifically, the model describes how, after some stimulation, the local blood flow $\boldsymbol{f}_{in}(t)$ increases and leads to the subsequent augmentation of the local capillary volume $\boldsymbol{\nu}(t)$. This incoming blood is strongly oxygenated but only part of the oxygen is consumed. It follows a local decrease of the deoxyhemoglobin concentration $\boldsymbol{\xi}(t)$ and therefore a BOLD signal variation. The Balloon model was then extended in [8] to include the effect of the neuronal activity $\boldsymbol{u}(t)$ on the variation of some auto-regulated flow inducing signal $\boldsymbol{\psi}(t)$ so as to eventually link neuronal to hemodynamic activity. The global physiological model corresponds then to a non-linear system with four state variables

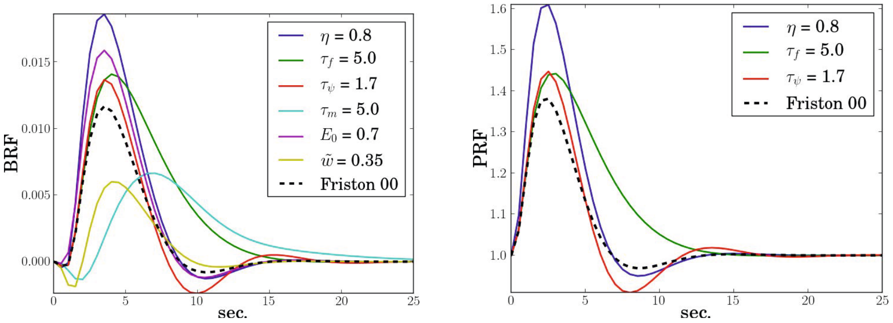


Fig. 1. Effect of the physiological parameters on the BRF (left) and PRF (right) shapes. The parameters values proposed in [8] are used except for one parameter whose identity and value is modified as indicated in the plot.

$\{\boldsymbol{\psi}, \boldsymbol{f}_{in}, \boldsymbol{\nu}, \boldsymbol{\xi}\}$ corresponding to normalized flow inducing signal, local blood flow, local capillary volume, and deoxyhemoglobin concentration. Their interactions over time are described by the following system of differential equations:

$$\begin{cases} \frac{d\boldsymbol{f}_{in}(t)}{dt} = \boldsymbol{\psi}(t) \\ \frac{d\boldsymbol{\psi}(t)}{dt} = \eta u(t) - \frac{\boldsymbol{\psi}(t)}{\tau_\psi} - \frac{\boldsymbol{f}_{in}(t)-1}{\tau_f} \\ \frac{d\boldsymbol{\xi}(t)}{dt} = \frac{1}{\tau_m}\left(\boldsymbol{f}_{in}(t)\frac{1-(1-E_0)^{1/\boldsymbol{f}_{in}(t)}}{E_0} - \boldsymbol{\xi}(t)\boldsymbol{\nu}(t)^{\frac{1}{\tilde{w}}-1}\right) \\ \frac{d\boldsymbol{\nu}(t)}{dt} = \frac{1}{\tau_m}\left(\boldsymbol{f}_{in}(t) - \boldsymbol{\nu}(t)^{\frac{1}{\tilde{w}}}\right) \end{cases} \tag{1}$$

with initial conditions $\boldsymbol{\psi}(0) = 0, \boldsymbol{f}_{in}(0) = \boldsymbol{\nu}(0) = \boldsymbol{\xi}(0) = 1$. Lower case notation is used for normalized functions by convention. The system depends on 5 hemodynamic parameters: τ_ψ, τ_f and τ_m are time constants respectively for signal decay/elimination, auto-regulatory feedback from blood flow and mean transit time, $\tilde{w}$ reflects the ability of the vein to eject blood, and E_0 is the oxygen extraction fraction. Another parameter η is the neuronal efficacy weighting term that models neuronal efficacy variability.

Once the solution of the previous system is found, Buxton et al [10] proposed the following expression that links the BOLD response $\boldsymbol{h}(t)$ to the physiological quantities considering intra-vascular and extra-vascular components:

$$\boldsymbol{h}(t) = V_0[k_1(1-\boldsymbol{\xi}(t)) + k_2(1-\frac{\boldsymbol{\xi}(t)}{\boldsymbol{\nu}(t)}) + k_3(1-\boldsymbol{\nu}(t))] \tag{2}$$

where k_1, k_2 and k_3 are scanner-dependent constants and V_0 is the resting blood volume fraction. According to [10],$k_1 \cong 7E_0$, $k_2 \cong 2$ and $k_3 \cong 2E_0 - 0.2$ at a field strength of 1.5T and echo time $TE = 40$ms.

The physiological parameters used are the ones proposed by Friston et al in [8]: $V_0 = 0.02$, $\tau_\psi = 1.25$, $\tau_f = 2.5$, $\tau_m = 1$, $\tilde{w} = 0.2$, $E_0 = 0.8$ and $\eta = 0.5$. The BRF and PRF generated using these parameters with the physiological model are shown in Fig. 1 under the label "Friston 00" (dashed line). The rest of the curves show the effect of changing the physiological parameters: η is a scaling factor and causes non-linearities above a certain value; τ_ψ controls the signal decay, which is more or less smooth; the auto-regulatory feedback τ_f regulates the undershoot; the transit time τ_m expands or contracts the signal in time; the windkessel parameter $\tilde{w}$ models the initial dip and the response magnitude; the oxygen extraction E_0 impacts the response scale. After analysing the behaviour of the model when varying the parameters values, the impact of each parameter was investigated and we concluded that the values proposed in [8] seemed reasonable.

2.2 Physiological Linear Relationship between Response Functions

From the system of equations previously defined, we derive an approximate relationship between the PRF, namely $\boldsymbol{g}(t)$, and the BRF, which is given by $\boldsymbol{h}(t)$ when $\boldsymbol{u}(t)$ is an impulse function. Both BRF and PRF are percent signal changes, and we consider $\boldsymbol{g}(t) = \boldsymbol{f}_{in}(t) - 1$, as $\boldsymbol{f}_{in}(t)$ is the normalized perfusion, with initial value 1. Therefore the state variables are $\{\boldsymbol{\psi}, \boldsymbol{g}, 1-\boldsymbol{\nu}, 1-\boldsymbol{\xi}\}$.

In the following we will drop the time index t and consider functions $\boldsymbol{h}, \boldsymbol{\psi}$, *etc.* in their discretized vector form. We can obtain a simple relationship between $\boldsymbol{h}$ and $\boldsymbol{g}$ by linearizing the system of equations. Equation (2) can first be linearized into:

$$\boldsymbol{h} = V_0[(k_1 + k_2)(1-\boldsymbol{\xi}) + (k_3 - k_2)(1-\boldsymbol{\nu})] \,. \tag{3}$$

We then linearize the system (1) around the resting point $\{\boldsymbol{\psi}, \boldsymbol{g}, 1-\boldsymbol{\nu}, 1-\boldsymbol{\xi}\} = \{\boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0}, \boldsymbol{0}\}$ as in [11]. From this linearization, denoting by $\mathcal{D}$ the first order differential operator and $\boldsymbol{I}$ the identity matrix, we get:

$$\begin{cases} \mathcal{D}\{\boldsymbol{g}\} = -\boldsymbol{\psi} \\ \left(\mathcal{D} + \frac{\boldsymbol{I}}{\tilde{w}\tau_m}\right)\{1-\boldsymbol{\nu}\} = -\frac{1}{\tau_m}\boldsymbol{g} \\ \left(\mathcal{D} + \frac{\boldsymbol{I}}{\tau_m}\right)\{1-\boldsymbol{\xi}\} = -\left(\gamma\boldsymbol{I} - \frac{1-\tilde{w}}{\tilde{w}\tau_m^2}\left(\mathcal{D} + \frac{\boldsymbol{I}}{\tilde{w}\tau_m}\right)^{-1}\right)\boldsymbol{g} \end{cases}, \tag{4}$$

where $\gamma = \frac{1}{\tau_m}\left(1 + \frac{(1-E_0)\ln(1-E_0)}{E_0}\right)$. It follows a linear link between $\boldsymbol{h}$ and $\boldsymbol{g}$ that we write as $\boldsymbol{g} = \boldsymbol{\Omega}\boldsymbol{h}$ where:

$$\boldsymbol{\Omega} = V_0^{-1}\left(-(k_1+k_2)\gamma\boldsymbol{B} + (k_1+k_2)\frac{1-\tilde{w}}{\tilde{w}\tau_m^2}\boldsymbol{B}\boldsymbol{A} - \frac{k_3-k_2}{\tau_m}\boldsymbol{A}\right)^{-1} \tag{5}$$

$$\text{with } \boldsymbol{A} = \left(\mathcal{D} + \frac{\boldsymbol{I}}{\tilde{w}\tau_m}\right)^{-1} \text{ and } \quad \boldsymbol{B} = \left(\mathcal{D} + \frac{\boldsymbol{I}}{\tau_m}\right)^{-1} \tag{6}$$

Using values of physiological constants as proposed in [8], Fig. 2 shows the BRF and PRF results that we get ($\boldsymbol{h}_{lin}$, $\boldsymbol{g}_{lin}$) by applying the linear operator to physiologically generated PRF ($\boldsymbol{g}_{physio}$) or BRF ($\boldsymbol{h}_{physio}$): $\boldsymbol{h}_{lin} = \boldsymbol{\Omega}^{-1}\boldsymbol{g}_{physio}$ or $\boldsymbol{g}_{lin} = \boldsymbol{\Omega}\boldsymbol{h}_{physio}$ compared to these physiologically generated $\boldsymbol{h}_{physio}$ and $\boldsymbol{g}_{physio}$

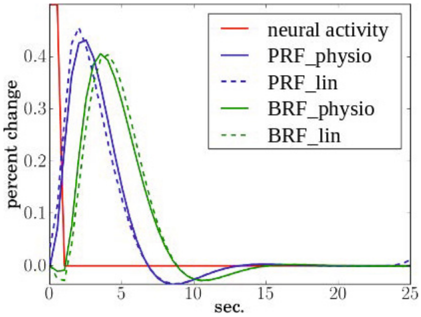


Fig. 2. Physiological responses generated with the physiological model, using parameters proposed in [8]: neural activity $\boldsymbol{\psi}$, physiological ($\boldsymbol{h}_{physio}$ or BRF$_{physio}$) and linearized ($\boldsymbol{h}_{lin}$ or BRF$_{lin}$) BRFs, physiological ($\boldsymbol{g}_{physio}$ or PRF$_{physio}$) and linearized ($\boldsymbol{g}_{lin}$ or PRF$_{lin}$) PRFs.

functions, computed by using the physiological model differential equations. Note that, although time-to-peak (TTP) values are not exact, the linear operator maintains the shape of the functions and satisfyingly captures the main features of the two responses. We considered a finer temporal resolution than TR for $\boldsymbol{\Omega}$ and, besides this, there is no direct dependence on the TR.

The derivation of this linear operator gives us a new tool for analyzing the ASL signal, although this link is subject to caution as linearity assumption is strong and this linearization induces approximation error.

3 Bayesian Hierarchical Model for ASL Data Analysis

The ASL JDE model described in [6,7] assumes a partitioned brain into several functional homogeneous parcels each of which gathers signals which share the same response shapes. In a given parcel $\mathcal{P}$, the generative model for ASL time series, measured at times $(t_n)_{n=1:N}$ where $t_n = n\text{TR}$, N is the number of scans and TR the time of repetition, with M experimental conditions, reads $\forall\, j \in \mathcal{P}$, $|\mathcal{P}| = J$:

$$\boldsymbol{y}_j = \sum_{m=1}^{M} \underbrace{a_j^m \boldsymbol{X}^m \boldsymbol{h}}_{(a)} + \underbrace{c_j^m \boldsymbol{W}\boldsymbol{X}^m \boldsymbol{g}}_{(b)} + \underbrace{\boldsymbol{P}\boldsymbol{\ell}_j}_{(c)} + \underbrace{\alpha_j \boldsymbol{w}}_{(d)} + \underbrace{\boldsymbol{b}_j}_{(e)} \tag{7}$$

The signal is decomposed into (a) task-related BOLD and (b) perfusion components given by the first two terms respectively; (c) a drift component $\boldsymbol{P}\boldsymbol{\ell}_j$ already considered in the BOLD JDE [5]; (d) a perfusion baseline term $\alpha_j \boldsymbol{w}$ which completes the modelling of the perfusion component; and (e) a noise term.

ASL fMRI data consists in the consecutive and alternated acquisitions of control and magnetically tagged images. The tagged image embodies a perfusion component besides the BOLD one, which is present in the control image too.

The BOLD component is noisier compared to standard BOLD fMRI acquisition. The control/tag effect is implicit in the ASL JDE model with the use of matrix $\boldsymbol{W}$. More specifically, we further describe each signal part below.

(a) The BOLD component: $\boldsymbol{h} \in \mathbb{R}^{F+1}$ represents the unknown BRF shape, with size $F+1$ and constant within $\mathcal{P}$. The magnitude of activation or BOLD response levels are $\boldsymbol{a} = \left\{a_j^m, j \in \mathcal{P}, m = 1:M\right\}$.
(b) The perfusion component: It represents the variation of the perfusion from the baseline when there is task-related activity. $\boldsymbol{g} \in \mathbb{R}^{F+1}$ represents the unknown PRF shape, with size $F+1$ and constant within $\mathcal{P}$. The magnitude of activation or perfusion response levels are $\boldsymbol{c} = \left\{c_j^m, j \in \mathcal{P}, m = 1:M\right\}$. $\boldsymbol{W}$ models the control/tag effect in the perfusion component, and it is further explained below.
(a-b) Considering $\Delta t < TR$ the sampling period of $\boldsymbol{h}$ and $\boldsymbol{g}$, whose temporal resolution is assumed to be the same, $\boldsymbol{X} = \{x^{n-f\Delta t}, n = 1:N, f = 0:F\}$ is a binary matrix that encodes the lagged onset stimuli. In [6,7], BRF and PRF shapes follow prior Gaussian distributions $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, v_{\boldsymbol{h}} \boldsymbol{\Sigma}_{\boldsymbol{h}})$ and $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, v_{\boldsymbol{g}} \boldsymbol{\Sigma}_{\boldsymbol{g}})$, with covariance matrices $\boldsymbol{\Sigma}_h$ and $\boldsymbol{\Sigma}_g$ encoding a constraint on the second order derivative so as to account for temporal smoothness. The BOLD (BRLs) and perfusion (PRLs) response levels (resp. $\boldsymbol{a}$ and $\boldsymbol{c}$) are assumed to follow different spatial Gaussian mixture models but governed by common binary hidden Markov random fields $\{q_j^m, j \in \mathcal{P}\}$ encoding voxels' activation ($q_j^m = 1, 0$ for activated, resp. non-activated) states for each experimental condition m. This way, BRLs and PRLs are independent conditionally to $\boldsymbol{q}$: $p(\boldsymbol{a}, \boldsymbol{c} \,|\, \boldsymbol{q})$. An Ising model on $\boldsymbol{q}$ introduces spatial correlation as in [6,7]. For further interest please refer to [5]. Univariate Gamma/Gaussian mixtures were used instead in [12] at the expense of computational cost. The introduction of spatial modelling through hidden Markov random fields gave an improved sensitivity/specificity compromise.
(c) The drift term: It allows to account for a potential drift and any other nuisance effect (e.g. slow motion parameters). Matrix $\boldsymbol{P} = \left[\boldsymbol{p}_1, \ldots, \boldsymbol{p}_O\right]$ of size $N \times O$ comprises the values of an orthonormal basis (*i.e.,* $\boldsymbol{P}^{\mathrm{t}}\boldsymbol{P} = \boldsymbol{I}_O$). Vector $\boldsymbol{\ell}_j = (\ell_{o,j}, o = 1:O)^{\mathrm{t}}$ contains the corresponding unknown regression coefficients for voxel j. The prior reads $\boldsymbol{\ell}_j \sim \mathcal{N}(0, v_{\boldsymbol{\ell}} I_O)$.
(b-d) The control/tag vector $\boldsymbol{w}$ (N-dimensional): It encodes the difference in magnetization signs between control and tagged ASL volumes. $w_{t_n} = 1/2$ if t_n is even (control volume) and $w_{t_n} = -1/2$ otherwise (tagged volume), and $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{w})$ is the diagonal matrix with $\boldsymbol{w}$ as diagonal entries.
(d) The perfusion baseline: It is encoded by α_j at voxel j. The prior reads $\alpha_j \sim \mathcal{N}(0, v_\alpha)$.
(e) The noise term: It is assumed white Gaussian with unknown variance v_b, $\boldsymbol{b}_j \sim \mathcal{N}(\boldsymbol{0}, v_b \boldsymbol{I}_N)$.
Hyper-parameters $\boldsymbol{\Theta}$. Non-informative Jeffrey priors are adopted for $\left\{v_{\boldsymbol{b}}, v_{\boldsymbol{\ell}}, v_{\boldsymbol{\alpha}}\right\}$ and proper conjugate priors are considered for the mixture parameters of BRLs ($\boldsymbol{\theta_a}$) and PRLs ($\boldsymbol{\theta_c}$).

4 A Physiologically Informed 2-steps Inference Procedure

The BOLD component is known to have a higher SNR than the perfusion component in the ASL signal, and can be estimated with a higher confidence. The link $\boldsymbol{g} = \boldsymbol{\Omega h}$ that we derived between both components can then be used to inform the PRF from the BRF. Using this link the other way around may not be satisfying as it may result in a contamination of $\boldsymbol{h}$ by a noisier $\boldsymbol{g}$.

This effect has been noticed in the implementation of a physiologically informed Bayesian procedure, considering the generative model (7), and the following priors for the BRF and PRF $\boldsymbol{h} \sim \mathcal{N}(0, v_{\boldsymbol{h}} \boldsymbol{\Sigma}_h)$ and $\boldsymbol{g}|\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{\Omega h}, v_{\boldsymbol{g}} \boldsymbol{\Sigma}_g)$, with $\boldsymbol{\Sigma}_h = \boldsymbol{\Sigma}_g = (\Delta t)^4 (\boldsymbol{D}_2^t \boldsymbol{D}_2)^{-1}$. $\boldsymbol{D}_2$ is the truncated second order finite difference matrix of size $(F-1) \times (F-1)$ that introduces temporal smoothness, as in [6,7], and $v_{\boldsymbol{h}}$ and $v_{\boldsymbol{g}}$ are scalars that we set manually. As seen in Fig. 4[Middle], this approach does not yield satisfying results, not only for the perfusion component, but also for the BOLD one, compared to the model presented in [6,7].

We therefore propose to exploit the described physiological link in a two-step procedure, in which we first identify hemodynamics properties ($\hat{\boldsymbol{h}}$, $\hat{a}_j^m$), and then use the linear operator $\boldsymbol{\Omega}$ and the previously estimated hemodynamic properties to recover the perfusion component ($\hat{\boldsymbol{g}}$, $\hat{c}_j^m$). This way, we avoid an arising contaminating effect of $\boldsymbol{g}$ on the estimation of $\boldsymbol{h}$, as in the one-step approach in Fig. 4[Middle]. Each step is based on a Gibbs sampling procedure as in [6,7].

4.1 Hemodynamics Estimation Step $\mathcal{M}_1$

In a first step $\mathcal{M}_1$, our goal is to extract the hemodynamic components and the drift term from the ASL data. In the JDE framework (7), it amounts to initially considering the perfusion component as a generalized perfusion term, including perfusion baseline and event-related perfusion response. The generative model (7) for ASL time series can be equivalently written, by grouping the perfusion terms involving $\boldsymbol{W} = diag(\boldsymbol{w})$, as

$$\boldsymbol{y}_j = \sum_{m=1}^{M} a_j^m \boldsymbol{X}^m \boldsymbol{h} + \boldsymbol{P\ell}_j + \boldsymbol{W} \left(\sum_{m=1}^{M} c_j^m \boldsymbol{X}^m \boldsymbol{g} + \alpha_j \mathbf{1} \right) + \boldsymbol{b}_j \tag{8}$$

where we consider $\alpha_j \boldsymbol{w} = \boldsymbol{W} \alpha_j \mathbf{1}$. Note that the hemodynamics components BRF $\boldsymbol{h}$ and the drift term $\boldsymbol{\ell}_j$ can be estimated first, by segregating them from a general perfusion term and a noise term. However, the perfusion component is considered in the residuals, so as to properly estimate its different contributions in a second step $\mathcal{M}_2$.

Given the estimated $\hat{\boldsymbol{h}}^{\mathcal{M}_1}$, $\hat{\boldsymbol{\ell}}^{\mathcal{M}_1}$ and $\hat{a}^{\mathcal{M}_1}$, we then compute residuals $\boldsymbol{r}^{\mathcal{M}_1}$ containing the remaining perfusion component:

$$\boldsymbol{r}_j^{\mathcal{M}_1} = \boldsymbol{y}_j - \sum_{m=1}^{M} \hat{a}_j^{m,\mathcal{M}_1} \boldsymbol{X}^m \hat{\boldsymbol{h}}^{\mathcal{M}_1} - \boldsymbol{P} \hat{\boldsymbol{\ell}}_j^{\mathcal{M}_1} \tag{9}$$

4.2 Perfusion Response Estimation Step $\boldsymbol{\mathcal{M}_2}$

From the residuals of the first step $\boldsymbol{r}^{\mathcal{M}_1}$, we estimate the perfusion component. The remaining signal is, according to (7), $\forall j = 1 : J$,

$$\boldsymbol{y}_j^{\mathcal{M}_2} = \boldsymbol{r}_j^{\mathcal{M}_1} = \sum_{m=1}^{M} c_j^m \boldsymbol{W}\boldsymbol{X}^m \boldsymbol{g} + \alpha_j \boldsymbol{w} + \boldsymbol{b}_j \qquad (10)$$

In this step, we introduce a prior on $\boldsymbol{g}$, to account for the already described physiological relationship $\boldsymbol{g} = \boldsymbol{\Omega}\boldsymbol{h}$:

$$\boldsymbol{g}|\widehat{\boldsymbol{h}}^{\mathcal{M}_1} \sim \mathcal{N}(\boldsymbol{\Omega}\widehat{\boldsymbol{h}}^{\mathcal{M}_1}, v_{\boldsymbol{g}}\boldsymbol{\Sigma}_g), \text{ with } \boldsymbol{\Sigma}_g = \boldsymbol{I}_F \ . \qquad (11)$$

The significance of the 2-step approach is to first preprocess the data to subtract the hemodynamic component within the ASL signal, as well as the drift effect, and to focus in a second step on the analysis of the smaller perfusion effect. In [4], differencing methods were used to subtract components with no interest in the perfusion analysis and directly analyse the perfusion effect in the time series. In contrast to these methods, we expect to disentangle perfusion from BOLD components by identifying all the components contained in the signal, and to recover them more accurately.

5 Simulation Results

The generative model for ASL time series in section 3 has been used to generate artificial ASL data. A low SNR has been considered, with $TR = 1$ s, mean $ISI = 5.03$ s, duration 25 s, $N = 325$ scans and two experimental conditions ($M = 2$) represented with 20×20-voxel binary activation label maps corresponding to BRL and PRL maps shown in Fig. 3. For both conditions: $(a_j^m|q_j = 1) \sim \mathcal{N}(2.2, 0.3)$ and $(c_j^m|q_j = 1) \sim \mathcal{N}(0.48, 0.1)$. Parameters were chosen to simulate a typical low SNR ASL scenario, in which the perfusion component is much lower than the hemodynamics component. A drift $\boldsymbol{\ell}_j \sim \mathcal{N}(0, 10\boldsymbol{I}_4)$ and noise variance $v_b = 7$ were considered. BRF and PRF shapes were simulated with the physiological model, using the physiological parameters used in [8].

In a low SNR context, the PRF estimate retrieved by the former approach developed in [6,7] is no physilogically relevant as shown in Fig. 4[(c), Top]. In the case of a physiologically informed Bayesian approach, considering a single-step solution as in Fig. 4[Middle], the perfusion component estimation is worse than

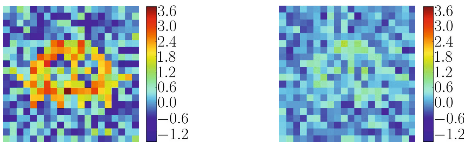


Fig. 3. BRL and PRL ground truth for a noise variance $v_b = 7$

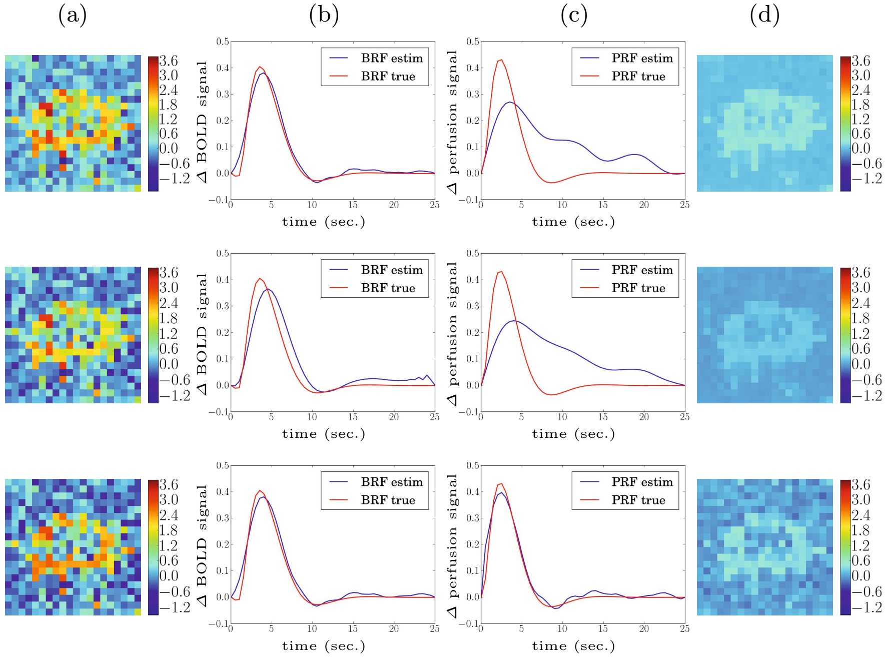


Fig. 4. Results on artificial data. **Top row**: non-physiological version. **Middle row**: physiological 1-step version. **Bottom row**: physiological 2-steps version. **(a,d)**: estimated BRL and PRL effect size maps respectively. The ground-truth maps for the BRL and PRL are depicted in Fig.3. **(b,c)**: BRF and PRF estimates, respectively, with their ground truth.

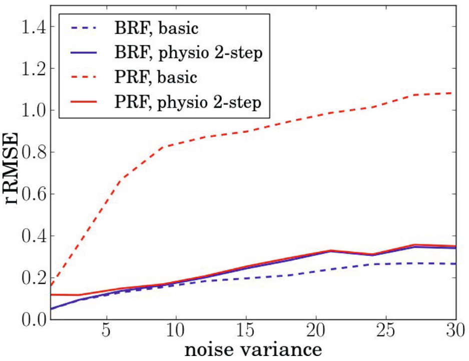


Fig. 5. Relative RMSE for the BRF and PRF and the two JDE versions, wrt noise variance v_b ranging from 0.5 to 30.

for the approach described in [6,7] and the BRF estimation is also degraded owing to the influence of the noisier perfusion component during the sampling. In contrast, the 2-steps method proposed here delivers a PRF estimate very close to the simulated ground truth (see Fig. 4[(c), Bottom] with a BRF which is well estimated too.

In Fig. 5, the robustness of both approaches with respect to the noise variance is studied, in terms of BRF and PRF recovery. The relative root-mean-square-error (rRMSE) is computed for the PRF and BRF estimates, i.e. $\text{rRMSE}_{\boldsymbol{\phi}} = \|\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}^{(true)}\| / \|\boldsymbol{\phi}^{(true)}\|$ where $\boldsymbol{\phi} \in \{\boldsymbol{h}, \boldsymbol{g}\}$. We observed that maintaining a good performance in the BRF estimation, we achieved a much better recovery of the PRF for noise variances larger than $v_b = 1$. Therefore, with the introduction of the physiological link between BRF and PRF, we have improved the PRF estimation.

6 Real Data Results

Real ASL data were recorded during an experiment designed to map auditory and visual brain functions, which consisted of $N = 291$ scans lasting $TR = 3$ s, with $TE = 18$ ms, FoV 192 mm, each yielding a 3-D volume composed of $64 \times 64 \times 22$ voxels (resolution of $3 \times 3 \times 3.5$ mm^3). The tagging scheme used was PICORE Q2T, with $TI_1 = 700$ ms, $TI_2 = 1700$ ms. The paradigm was a fast event-related design (mean $ISI = 5.1$ s) comprising sixty auditory and visual stimuli. Two regions of interest in the right temporal lobe, for the auditory cortex, and left occipital lobe, for the visual cortex, were defined manually.

Fig. 6(b-c) depicts the response estimates superimposed to the canonical shape which is in accordance with the BRF estimates for both methods. Indeed, we consider here an auditory region where the canonical version has been fitted. Accordingly, the BRL maps (Fig. 6(a)) also look alike for both methods. However, PRF estimates significantly differ and the effect of the physiologically-inspired regularization yields a more plausible PRF shape for the 2-steps approach compared with the non-physiological JDE version. Results on PRL maps (Fig. 6(d)) confirm the improved sensitivity of detection for the proposed approach. In the same way, in the visual cortex, Fig. 6(f-g) shows the BRF and PRF estimates, giving a more plausible PRF shape for the 2-steps approach, too. For the detection results (Fig. 6(h)), the 2-steps approach seems also to provide a much better sensitivity of detection.

7 Discussion and Conclusion

Starting from non-linear systems of differential equations induced by physiological models of the neuro-vascular coupling, we derived a tractable linear operator linking the perfusion and BOLD responses. This operator showed good approximation performance and demonstrated its ability to capture both realistic perfusion and BOLD components. In addition, this derived linear operator was easily incorporated in a JDE framework at no additional cost and with a significant improvement in PRF estimation, especially in critical low SNR situations. As shown on simulated data, the PRF estimation has been improved while maintaining accurate BRF estimation. Real data results seem to confirm the better performance of the proposed physiological approach compared to its competing alternative. In terms of validation, future work will be devoted to intensive validation on whole brain analysis and multiple subjects.

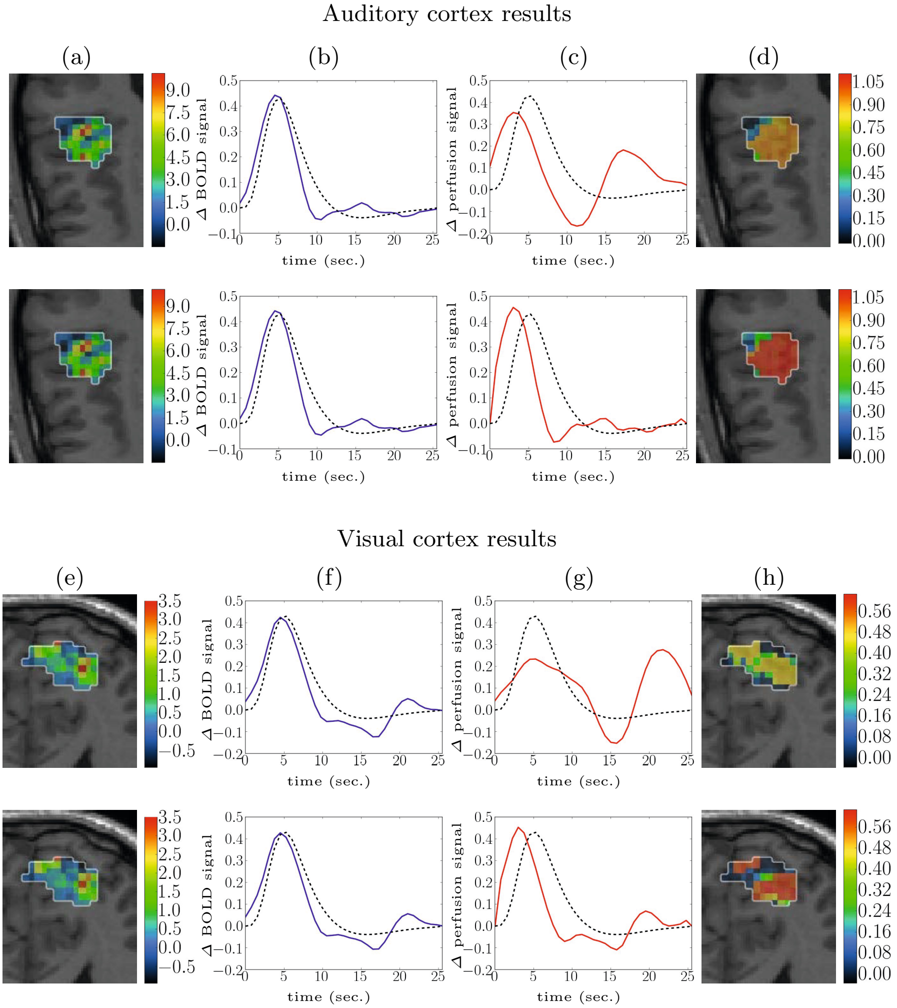


Fig. 6. Comparison of the two JDE versions on real data in the auditory and visual cortex. **(top row in auditory and visual cortex results)**: non-physiological version. **(bottom row in auditory and visual cortex results)**: physiological 2-steps version. **(a,e)** and **(d,h)**: estimated BRL and PRL effect size maps, respectively. **(b,f)** and **(c,g)**: BRF and PRF estimates, respectively. The canonical BRF is depicted as a black dashed line, while PRF and BRF estimated are depicted in solid red and blue lines, respectively.

References

1. Williams, D., Detre, J., Leigh, J., Koretsky, A.: Magnetic resonance imaging of perfusion using spin inversion of arterial water. Proceedings of the National Academy of Sciences 89(1), 212–216 (1992)

2. Ogawa, S., Tank, D., Menon, R., Ellermann, J., Kim, S.G., Merkle, H., Ugurbil, K.: Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. Proceedings of the National Academy of Sciences 89, 5951–5955 (1992)
3. Hernandez-Garcia, L., Jahanian, H., Rowe, D.B.: Quantitative analysis of arterial spin labeling fmri data using a general linear model. Magnetic Resonance Imaging 28(7), 919–927 (2010)
4. Mumford, J.A., Hernandez-Garcia, L., Lee, G.R., Nichols, T.E.: Estimation efficiency and statistical power in arterial spin labeling fmri. Neuroimage 33(1), 103–114 (2006)
5. Vincent, T., Risser, L., Ciuciu, P.: Spatially adaptive mixture modeling for analysis of within-subject fMRI time series. IEEE Transactions on Medical Imaging 29(4), 1059–1074 (2010)
6. Vincent, T., Warnking, J., Villien, M., Krainik, A., Ciuciu, P., Forbes, F.: Bayesian Joint Detection-Estimation of Cerebral Vasoreactivity from ASL fMRI Data. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 616–624. Springer, Heidelberg (2013)
7. Vincent, T., Forbes, F., Ciuciu, P.: Bayesian BOLD and perfusion source separation and deconvolution from functional ASL imaging. In: 38th Proc. IEEE ICASSP, Vancouver, Canada, pp. 1003–1007 (May 2013)
8. Friston, K.J., Mechelli, A., Turner, R., Price, C.J.: Nonlinear responses in fMRI: The balloon model, Volterra kernels, and other hemodynamics. Neuroimage 12, 466–477 (2000)
9. Buxton, R.B., Uludağ, K., Dubowitz, D.J., Liu, T.T.: Modeling the hemodynamic response to brain activation. Neuroimage 23, 220 (2004)
10. Buxton, R.B., Wong, E.C., Frank, L.R.: Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. Magnetic Resonance in Medicine 39, 855–864 (1998)
11. Khalidov, I., Fadili, J., Lazeyras, F., Van De Ville, D., Unser, M.: Activelets: Wavelets for sparse representation of hemodynamic responses. Signal Processing 91(12), 2810–2821 (2011)
12. Makni, S., Ciuciu, P., Idier, J., Poline, J.B.: Bayesian joint detection-estimation of brain activity using MCMC with a Gamma-Gaussian mixture prior model. In: 31th Proc. IEEE ICASSP, Toulouse, France, vol. V, pp. 1093–1096 (May 2006)

Bone Reposition Planning for Corrective Surgery Using Statistical Shape Model: Assessment of Differential Geometrical Features

Neda Sepasian[1], Martijn Van de Giessen[2], Iwan Dobbe[3], and Geert Streekstra[3]

[1] Department of Biomedical Image Analysis, TU/e, Eindhoven, The Netherlands
[2] Division of Image Processing, LUMC, Leiden, The Netherlands,
[3] Department of Biomedical Engineering and Physics, AMC,
Amsterdam, The Netherlands

Abstract. We discuss a new planning method for corrective osteotomy surgery without the need to make a CT scan of the contralateral bone. We use a statistical shape model to estimate the most likely relative position of two bone segments of an osteotomized bone. To investigate the added value of geometrical properties for planning, different geometrical features of the bone surface are being incorporated. The feasibility and accuracy of our proposed method are investigated using 10 virtually deformed radii and a statistical shape model based on 35 healthy radii.

1 Introduction

Limb fractures are very common and sometimes result in malunion of the fractured bone segments causing chronic pain, reduced function and finally osteoarthritis. For the distal radius, e.g., the annual incidence rate is approximately 0.3% of the population each year [1], while about 5% of these cases result in a symptomatic malunion requiring secondary treatment by corrective osteotomy surgery [2]. In this procedure the bone is cut in two segments which are repositioned and fixated, mostly using an anatomical plate and screws.

In state-of-the-art techniques for planning of such a surgical reconstruction virtual 3-D bone models are created from CT volume data of the affected bone. In this approach the CT data of the mirrored contralateral limb is used as a reconstruction target in the planning procedure. [3–6]. A drawback of this approach is that a healthy contralateral reference is not always available. Also, this standard approach requires a CT scan of the healthy limb as well, which increases the radiation dose by a factor of two.

To overcome these drawbacks a method is required that can provide the planning of the surgical reconstruction based on the affected bone only by using shape information of the unaffected segments of the bone. For such a planning method application of a statistical shape model (SSM) of the bone [7, 8] fitted to the surface data of the unaffected bone parts seems a logical choice which allows prediction of the optimal alignment of the bone segments after surgery. In this context the SSM model describes the patterns that exist within the variation in shape in a population [9].

M.J. Cardoso et al. (Eds.): BAMBI 2014, LNCS 8677, pp. 49–60, 2014.

Recently SSM's were reported to be advantageous in a large number of orthopedic applications such as robust and fast bone segmentation [10, 11], creation of bone geometries for finite element modeling [12, 13], implant design optimization [14, 15] and several diagnostic applications [9, 16]. Related to our problem SSM's have been used to reconstruct bone surfaces from incomplete bone geometry representations for surgical planning and navigation [17]. In this paper we make an essential additional step by incorporating the alignment of the bone segments in the fitting procedure of the SSM to the incomplete bone surface data.

The general framework concerning the surface fitting problem use deviations in surface geometry between a reference shape and a target shape to fit one single shape to the other [18, 19]. The classic method for surface fitting is based on the iterative closest point (ICP) algorithm. This method minimizes the distance between points in one surface and the closest points in another surface [20, 21]. A surface matching algorithm was also developed for fitting a statistical shape model (SSM) of a same type of surface to a target surface [22]. In these methods, random point sets on a surface are being identified and these point clouds are registered and matched to compute the corresponding points.

A common way to control surface fitting is to use distance information between the reference and target during optimization, e.g., using Euclidean norms. This choice however exploits only limited information about the surface geometry. It is possible however, to add additional geometrical features such as different shape related vector fields and curvature as ingredients in controlling the fit. To this end, different approaches have been developed for matching source and target surface invariants such as curvature maps [19, 23]. All these methods require the iteratively updating of corresponding points during the fitting procedure which is a time consuming step. Furthermore, it is known that including the corresponding points results in a convergence to local minima due to the partial alignment instead of the global minimum indicating the perfect alignment.

In this paper we propose a method for surface-fitting of a SSM to the geometrical representation of an affected bone to plan realignment of two bone segments during corrective surgery. Moreover, we investigate the efficiency of using different metrics for fitting the SSM to the segments of a target bone. To this end we optimize the fitting of the model to the target bone using individual features such as spatial distances, curvature and curvature vectors. Primarily, our goal is to develop a method which does not require the availability of corresponding points, nor the data set of the contralateral bone. However, this requirement becomes essential in the presence of the curvature and curvature vectors. To this end, we propose a cost function that can be customized for this purpose. The residual positioning error for different optimization metrics is examined experimentally, using virtual malunions of radii outside those used for building the SSM.

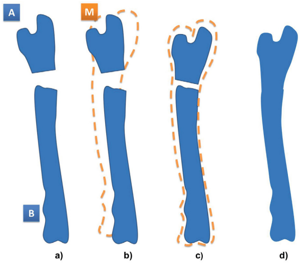


Fig. 1. Sketch for illustrating the correction prototype, a) Initial bone segment positions, b) Initial bone position M c) Intermediate step where the model is deformed, scaled and translated and bone segment A has been translated and rotated, d) the final model where the most likely solution has been found

2 SSM Based Planning

The challenge is to find the correct relative position of the distal and proximal bone segments from a single malunited bone. The surfaces A and B of both bone segments are initially sub-optimally aligned due to the malunion; see Figure 1.a. To find the correct alignment of these bone segments we propose to fit a statistical shape model M built from multiple segmentations of bones to patient data containing the two segments A and B of the bone. For initialization we use the statistical model M close to the two bone segments (Figure 1.b). Subsequently, an iterative optimization process is performed in which A is subject to translations and rotations, B is considered fixed, and M is translated, rotated, scaled and reshaped until A and B optimally fit with M. After convergence the translation and rotation parameters of A with respect to B describe the repositioning parameters that needs to be applied to A during surgery (Figure 1.c). The best-fitting shape M describes the most likely shape of the original bone (Figure 1.d).

An extra-articular fracture resulting in a malunion often shows a deformed region between a distal (A) and proximal (B) bone segment. In a malunion the distal bone segment is malpositioned with respect to the proximal bone segment, but apart from the location of malunion, the shapes of these bone segments are unaffected. Since the deformed region is unlike normal bone geometry we exclude it from the fitting procedure in our planning method.

2.1 Fitting of the SSM to Two Bone Segements

During the fitting process bone segment A and the model M are subject to translation t_A and t_M and rotation r_A and r_M computed by three Rodrigues rotation parameters. The model M is also allowed to scale indicated by the parameter s_M and to distort using the shape parameter $\boldsymbol{b}$. The variables to be optimized are $\boldsymbol{R} = [r_A; t_A; r_M; t_M; s_M; \boldsymbol{b}] = [\boldsymbol{R}_A; \boldsymbol{R}_M; \boldsymbol{b}]$.

The total likelihood to correct the two bone parts A and B orientation and position with respect to each other using a model M consists of the shape similarity, scaling and feature measures and reads

$$L(A, B, M, \boldsymbol{R}) = P_f(B \cup A(\boldsymbol{R}_A) \mid M(\boldsymbol{R}_M)) + P_s(s_M) + P_b(\boldsymbol{b}). \quad (1)$$

where $P(\boldsymbol{b})$ is a probability density function representing the validity of a shape with shape parameter $\boldsymbol{b}$, $P(s_M)$ is a probability function for scaling and $P_f(X \mid M)$ is a probability density function for measuring the similarity between the statistical model M and bone surface X. For our particular application we have $X = B \cup A(\boldsymbol{R}_A)$. The probability density function V includes the representation of surfaces geometrical features, shape deformation and the closet neighborhood distance measure.

The optimal composition of bone parts A and B can then be obtained by maximizing $L(A, B, M, \boldsymbol{R})$. However, due to very low likelihood values in $P_f(B \cup A(\boldsymbol{R}_A) \mid M(\boldsymbol{R}_M))$, this might lead to numerical problems. Therefore we minimize the negative logarithm $-\log P_f(B \cup A(\boldsymbol{R}_A) \mid M(\boldsymbol{R}_M))$ using a standard gradient descent method. In the following, building of the SSM as well as each of the probability density functions are described in detail.

2.2 Probability Distribution Functions for Shape Validity and Scaling

In order to construct the SSMs in this paper the active shape modelling introduced in [7] has been applied. We represent each bone by a $3n$ element vector formed by concatenating the elements of the individual surface points $\boldsymbol{x}_i = [x_1, y_1, z_1, \cdots, x_n, y_n, z_n]$, $i = 1, 2, \cdots, l$ where l is the number of individual shapes. Formally, each shape can be described using the linear model

$$\boldsymbol{x} = \boldsymbol{m} + \boldsymbol{P}\boldsymbol{b}. \quad (2)$$

Here, $\boldsymbol{m}$ consists of the coordinates of the mean shape, $\boldsymbol{P}$ is a matrix with modes of variation and $\boldsymbol{b}$ is a vector with the weighting parameters for the variations specified for each mode j. The non-rigid registration introduced in [24] has been applied in order to estimate the corresponding points of all shapes and consequently the mean $\boldsymbol{m}$ and the modes of variation $\boldsymbol{P}$ [7]. Using the weighted summation of the different modes of variations, a new shape can be computed.

The probability distribution function for scaling $P_s(s_M)$ is modeled using a 1D normal distribution with mean 1 and standard deviation s_M computed from the volume estimation during the SSM construction.

2.3 Probability Distribution Model to Compare Shapes

We propose a probability distribution function inspired by the registration model introduced in Granger *et al.* [25]. In this model the alignment of two point clouds is being treated as a probability density maximization problem, where one point clouds is representing the centroid of a Gaussian Mixture Model (GMM) and the other one represents the data points. Ideally, two point sets become aligned and the correspondence is estimated using the Mahalanobis distance. Here, bone segment A are described by points a_j on the surface of A and points m_k are located on the surface of the deformed and transformed model M is described by a GMM. We propose a probability distribution which combines the point-wise distance (e.g. Euclidean) between the model and the patient data, the point-wise angle between any vector data corresponding to the point cloud shapes, and the differences in their curvature maps. Given the points $a_j; j = 1; 2; \cdots ; n_A$ on A with n_A as the number of points in cloud A, then the likelihood that the point m_k in M is sampled as point on A is computed by

$$P(a_j \mid m_k) = \frac{1}{(2\pi)^{9/2} \left(\mid \Sigma_p \mid\mid \Upsilon_p \mid\mid \Gamma_p \mid\right)^{3/2}} \tag{3}$$

$$\exp(-\frac{1}{2}(\boldsymbol{d}_{jk}^{\top} \Sigma_p^{-1} \boldsymbol{d}_{jk})) + \exp(-\frac{1}{2}(\boldsymbol{c}_{jk}^{\top} \Upsilon_p^{-1} \boldsymbol{c}_{jk})) + \exp(-\frac{1}{2}(\boldsymbol{t}_{jk}^{\top} \Gamma_p^{-1} \boldsymbol{t}_{jk}))$$

Where the covariance matrix $\Sigma_p = \sigma_p I$ is a diagonal 3×3 matrix with the standard deviation σ_p and the identity matrix I. Respectively, $\Upsilon_p = \epsilon_p I$ and $\Gamma_p = \gamma_p I$ describe the different Covariance matrices for curvature and the vectors. In this work all n_M points of m_k; $k = 1; 2; \cdots ; n_M$ in M are considered equally uncertain and therefore standard deviations are the same for all m_k. The Euclidean distance is $d_{jk}(a_j, m_k) = d_{jk} = \sqrt{\sum_{i=1}^{3}(a_j^i - m_k^i)}$ and the vector match measure is $t_{jk}(a_j, m_k) = t_{jk} = 1 - w(a_j, m_k)$. Where, $w(a_j, m_k) = v_j.v_k/|v_j|.|v_k|$, $0 \leq w \leq 1$ and v_j is the vector at point a_j where $|.|$ denotes the norm of the vector. The smaller is the angle between vectors at two points, the larger is the similarity between two point data. The vectors denote the principal curvature vectors. Later on in Section 5 we will clarify the principal curvature vector definition. Here, $c_{jk} = c(a_j, m_k)$ represents the difference between the curvature at a_j and m_k. Our probability function is customized in order to include the corresponding points. This setting possible as long as the nature of distribution of all the features can be well defined.

Given the statistical model (3), the likelihood $P(X \mid M)$ that all points in X are sampled from M is estimated by

$$P_f(X \mid M) = \sum_{j=1}^{n_A} P(a_j \mid M) = \sum_{j=1}^{n_A} \frac{1}{n_M} \sum_{k=1}^{n_M} P(a_j \mid m_k). \tag{4}$$

Where X is one or more bone segments.

3 Experiments

3.1 Data

The proposed method was evaluated using CT scans of healthy radii. To this end 45 radii (26 women, 19 men, age [11, 56] years) were imaged with voxel size

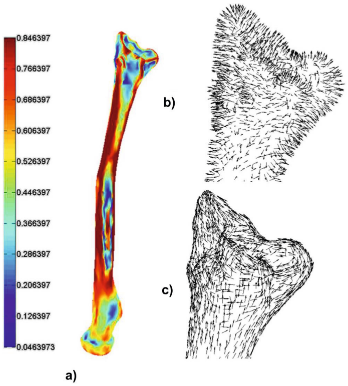


Fig. 2. a) Gaussian curvature map, the color bar indicates the maximum principal curvature, b) Normal vector illustration, c) maximum curvature vector

(0.45 × 0.45 × 0.45)mm using a Brilliance 64-channel CT scanner with a regular-dose, high-resolution protocol. In order to scan the complete radii individuals were scanned in prone position with the forearm extended above the head. Right radii polygons were created by image segmentation [4]. We mirror the cases where only the left radius was scanned. This resulted in 45 healthy right radii polygons. Ten of these were randomly selected as target bones, the remaining 35 bones were used as training shapes for building SSMs. Knowing the corresponding points of the bones inside the training set, enabled estimating σ, ϵ and γ using the standard deviation of the differences between the corresponding points distance-, curvature- and curvature vector-wise. Subsequent experiments were done using these values. In case more than one of these surface features was evaluated at the same time, linear combinations of the distance d_{jk}, curvature c_{jk} and tangent vector measure t_{jk} applied (3). We sampled 5000 points per surface. In the above-described experiment, optimization of the fitting procedure is based on the average nearest neighbor distance between points of the SSM and a target bone.

3.2 Evaluating the SSM

To investigate how many shapes are required to build a SSM that sufficiently represent a set of target bones, we randomly selected 10, 15, 20, 25 and 35 training shapes of complete bones to build the SSMs. This SSM was subsequently fit to the 10 target bones and the closet nearest neighbor distance was determined for each target bone, resulting in a mean and a standard deviation of this parameter. Figure 3.a shows the mean error and standard deviation values for these experiments. It is clearly shown that the mean and standard deviation reduce with the number of training shapes in the SSM. Our computations indicate the SSM containing 35 training shapes is considered sufficiently accurate to describe a target bone shape and is therefore used as SSM in the remainder of this document.

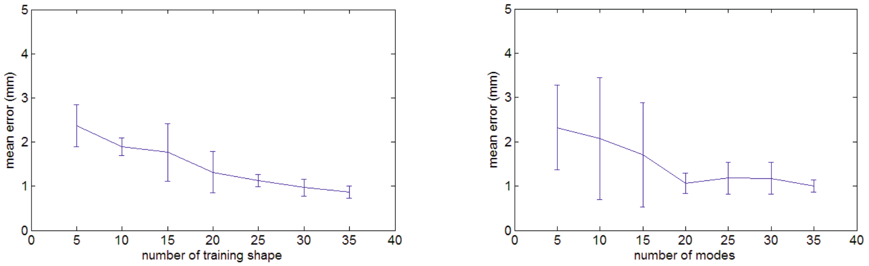


Fig. 3. a) Agreement between SSM and target bones as a function of the number of training shapes inside the SSM. b) Agreement between SSM, containing 35 training shapes, and target bones as a function of the number of modes of variation taken into account.

3.3 Evaluating Modes of Variation

The fitting procedure can be accelerated by only considering relevant modes of variation and disregarding higher modes of variation that merely represent noise. To evaluate how many modes of variation are required to fit the SSM with sufficient accuracy to a set of target bones, we performed fitting of the SSM to the 10 target bones taking into account an increasing number of modes of variation Fig. (3b). The nearest neighbor distance is again used to quantify the agreement between SSM and target bones. Based on this computations for the rest of the experiments we chose the first 25 modes of variations.

3.4 Including the Geometrical Features

We performed fitting of the SSM to the 10 target bones using different combinations of the geometrical features, 35 training shapes and 25 modes of variations. We estimating corresponding points based on minimum Euclidean distance. Figure 4 shows the computational result. Note that this figure only illustrates the result for the combination with more obvious variations with respect to the rest. We observe that including the geometrical features result in a more consistent fitting and smaller distribution of the standard deviation. This is due to the improvement of the local alignment. Using this computation for the repositioning experiment we choose the combination of distance and the curvature vector measure.

3.5 Accuracy of Bone Repositioning

To evaluate the accuracy of bone repositioning using the proposed method we simulated three virtual malunions for each of the 10 target bones, i.e., 30 simulated malunions. This enables comparing the reconstructed radius with its ground truth, i.e., the radius before the malunion. A malunion was simulated by

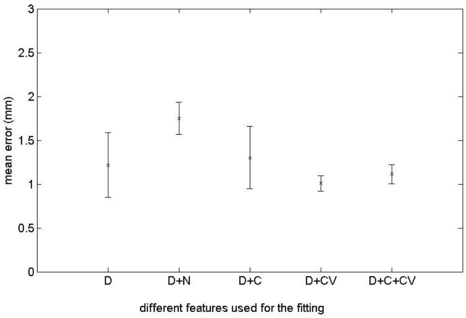


Fig. 4. Variation of the fitting using combination of different features using 35 training shapes and 25 modes of variations, D stands for distance, N is for normal vectors, C is for curvature, CV stands for curvature vectors.

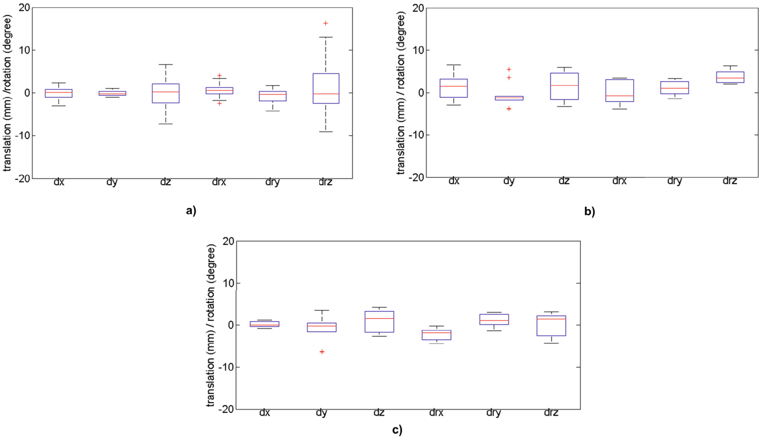


Fig. 5. Boxplot showing the accuracy of repositioning the distal radius segment with respect to the proximal segment, based on fitting a SSM to remaining bone segments in a simulation experiment (10 target bones, 3 deformities per bone). dx, dy and dz are translations between the centers of gravity of the reconstruction result and the ground truth. drx, dry and drz are the rotations about the x, y and z axes. a)reconstruction using bilateral asymmetry [26] b) reconstruction using distance measure. c) reconstruction using distance and curvature vectors.

removing a bone piece between a distal and proximal bone segment and by randomly translating (range $[2, 5]$mm) and rotating (range $[10, 60]^\circ$) the distal part. The height of the randomly removed piece (the defect) is called the defect height (range [3, 7] mm). The simulated bone was subsequently reconstructed using the proposed method. Figure 5 shows the residual rotation r_r and translation t_r error with respect to the ground truth as box plots. The box represents standard deviation , the horizontal line the mean value and the whiskers represent the range. Figure 5.a indicates differences due to bilateral asymmetry reported by Vroemen et al. [26] as determined for 20 individuals by left-to-right matching and right-to-left matching of distal and proximal segments, explaining the different sign in their results. These dots enable comparing repositioning results of the proposed method with generally accepted variations since the contralateral side is normally considered the best reference available. Figure 5a shows the estimated error when only the probability distance measure is used as metric during the fitting procedure.

We observe the improvement of the error distribution level in comparison with the fit using bilateral differences. The translation errors dx, dy and dz are smaller than acceptable levels considering the bilateral differences. For rotation, also we see an improvement in comparison with the bilateral differences particularly about the z-axis, i.e., the longitudinal axis of the bone. In general, we observe a slight improvement by adding curvature geometrical features; see Figure 5b and 5c. Figure 6b is a typical example after reconstruction with only distance as optimization metric. Figure 6c – d shows the common reconstruction error which occurs by only using the distance measure. The experiments were performed for different defect heights and we noticed the larger error occurs with larger heights.

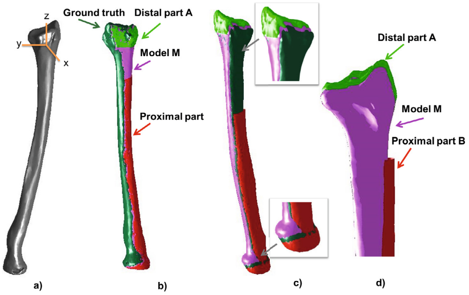


Fig. 6. a) Axes defined within the radius and heights of removals, b) correctly reconstructed bone, c) common translation and rotation error occurs using only the distance d) poorly reconstructed bone with 10 degrees rotation error about the z-axis and 4 mm translation error along the x-axis.

4 Discussion

We introduced a new technique for repositioning bone segments after a fracture or malunion using a SSM of the radius and a set of differential geometry features for surface-to-surface fitting. The method yields a set of transformation and rotation parameters, which can be used in a device for the actual bone repositioning; see [27]. We showed that residual positioning errors are very close to what can be achieved compared to standard 3-D planning, which is limited as a result of bilateral differences [26]. The use of only a distance measure as metric for the optimization of fitting seemed slightly inferior to using a metric that includes the shapes curvature and/or curvature vectors, particularly, when a larger piece of the bone is removed.

Our study aimed at modeling the radius although it can be easily extended for other bone types as well. The presented method requires an initial segmentation step. In future studies this can be avoided by performing the fitting and obtaining the geometrical features directly using CT grayscale volumes. It reduces a possible observer bias and increases the degree of automation since no user interaction is required for segmentation.

A big advantage of the proposed method is the fact that a contralateral bone is not required. This allows bone repositioning when a contralateral bone is not available. In addition, it reduces the radiation-absorbed dose, since the contralateral arm does not need to be scanned. In this pilot study residual errors in repositioning parameters already appeared to be very close to what can be achieved compared to conventional 3-D planning based on bilateral symmetry, we expect that the method will show valuable in the next generation of planning applications.

5 Appendix

Let $x_i = x_i(u, v), i = 1, 2, 3$ be a regular parameterizations of a surface. The Gaussian curvature of a surface in R^3 is given by

$$k = \kappa_1 \kappa_2 = \frac{LN - M^2}{EG - F^2},$$

where κ_1 and κ_2 are the principal curvature, $E = x_u \cdot x_u, F = x_u \cdot x_v$ and $G = x_v \cdot x_v$ are coefficients of the first fundamental form and $L = x_{uu} \cdot n, N = x_{vv} \cdot n, M = x_{uv} \cdot n_i$, are coefficients of the second fundamental form. These coefficients are computed at given point x_i in the parametric plane by the projections of the second partial derivatives of x at that point onto the normal vector n. In this setting, it is easy to see why the Gaussian curvature is independent of the choice of the unit normal n. Notice that if the sign of n is reversed, the signs of the coefficients of L, M, N are reversed too. Further, while the signs of both principal curvatures κ_1 and κ_2, the product $K = \kappa_1 \kappa_2$ remains unaffected. Clearly, the sign of mean curvature $H = \frac{(\kappa_1 + \kappa_2)}{2}$, depends on the choice of sign of n. In order to compute the curvature vectors, we compute the terms of the first and second fundamental forms and define the corresponding metric tensors [28],

$$F_1 = \begin{bmatrix} E & F \\ F & G \end{bmatrix}, \quad F_2 = \begin{bmatrix} L & M \\ M & N \end{bmatrix}, \tag{5}$$

and we introduce

$$O = inv(F_1)F_2.$$

The previously introduced principal curvatures κ_1, κ_2 are the eigenvalues of the matrix O. The eigenvectors of the matrix O are corresponding to the vectors pointing to the direction of κ_1, κ_2 respectively (Figure $6a - c$). Therefore, a shape can be described by the type of curvature and the type of the orientation.

References

1. Athwal, G.S., Ellis, R.E., Small, C.F., Pichora, D.R.: Computer-assisted distal radius osteotomy. The Journal of Hand Surgery 28(6), 951–958 (2003)
2. Miyake, J., Murase, T., Moritomo, H., Sugamoto, K., Yoshikawa, H.: Distal radius osteotomy with volar locking plates based on computer simulation. Clin. Ortho. and Rel. Res. 469(6), 1766–1773 (2011)
3. Cronier, P., Pietu, G., Dujardin, C., Bigorre, N., Ducellier, F., Gerard, R.: The concept of locking plates. Orthopaedics and Traumatology: Surgery and Research 96(4, Suppl.), S17–S36 (2010)
4. Dobbe, J., Strackee, S.D., Schreurs, A.W., Jonges, R., Carelsen, B., Vroemen, J., Grimbergen, C.A., Streekstra, G.J.: Computer-assisted planning and navigation for corrective distal radius osteotomy, based on pre- and intraoperative imaging. IEEE Trans. Biomed. Eng. 58(1), 182–190 (2011)
5. Dobbe, J., Vroemen, J.C., Strackee, S., Streekstra, G.: Patient-tailored plate for bone fixation and accurate 3d positioning in corrective osteotomy. Med. Biol. Eng. Comput. (2012) (in press)
6. Murase, T., Oka, K., Moritomo, H., Goto, A., Yoshikawa, H., Sugamoto, K.: Three-dimensional corrective osteotomy of malunited fractures of the upper extremity with use of a computer simulation system. The Journal of Bone and Joint Surgery 90(11), 2375–2389 (2008)
7. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models - their training and application. Comp. Vis. and Im. Und. 61(1), 38–59 (1995)
8. van de Giessen, M., Foumani, M., Vos, F.M., Strackee, S.D., Maas, M., Vliet, L.V., Grimbergen, C.A., Streekstra, G.J.: A 4d statistical model of wrist bone motion patterns. IEEE Trans. Med. Imaging 31(3), 613–625 (2012)
9. Waarsing, J., Rozendaal, R., Verhaar, J., Bierma-Zeinstra, S., Weinans, H.: A statistical model of shape and density of the proximal femur in relation to radiological and clinical OA of the hip. Osteoarthritis and Cartilage 18(6), 787–794 (2010)
10. Chung, F., Schmid, J., Magnenat-Thalmann, N., Delingette, H.: Comparison of statistical models performance in case of segmentation using a small amount of training datasets. The Visual Computer 27(2), 141–151 (2011)
11. Schmid, J., Kim, J., Magnenat-Thalmann, N.: Robust statistical shape models for mri bone segmentation in presence of small field of view. Med. Im. Analys., 155–168 (2011)
12. Nicolella, D.P., Bredbenner, T.L.: Development of a parametric finite element model of the proximal femur using statistical shape and density modelling. Computer Methods in Biomechanics and Biomedical Engineering 15(2), 101–110 (2012)
13. Taylor, M., Bryan, R., Galloway, F.: Accounting for patient variability in finite element analysis of the intact and implanted hip and knee: a review. Int. J. Numer. Method Biomed. Eng. 29(2), 273–292 (2013)
14. Kozic, N., Weber, S., Büchler, P., Lutz, C., Reimers, N., Ballester, M.G., Reyes, M.: Optimisation of orthopaedic implant design using statistical shape space analysis based on level sets. Medical Image Analysis 14(3), 265–275 (2010)
15. Mahfouz, M., Fatah, E.E.A., Bowers, L.S., Scuderi, G.: Three-dimensional morphology of the knee reveals ethnic differences. Clin. Orthop. Relat. Res. 407(1), 172–185 (2012)
16. Whitmarsh, T., Fritscher, K.D., Humbert, L., del Río Barquero, L.M., Roth, T., Kammerlander, C., Blauth, M., Schubert, R., Frangi, A.F.: Hip fracture discrimination from dual-energy x-ray absorptiometry by statistical model registration. Bone 51(5), 896–901 (2012)

17. Baek, S.Y., Wang, J.H., Song, I., Lee, K., Lee, J., Koo, S.: Automated bone landmarks prediction on the femur using anatomical deformation technique. Computer-Aided Design 45(2), 505–510 (2013)
18. Johnson, A.E., Hebert, M.: Surface matching for object recognition in complex 3-d scenes. Image and Vision Computing 16, 635–651 (1998)
19. Wang, Y., Peterson, B.S., Staib, L.H.: 3d brain surface matching based on geodesics and local geometry. Comp. Vis. and Im. Under. 89(2-3), 252–271 (2003)
20. Feldmar, J., Malandain, G., Declerck, J., Ayache, N.: Extension of the icp algorithm to non-rigid intensity-based registration of 3d volumes. In: Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, vol. 14(2), pp. 84–93 (1996)
21. Caunce, A., Taylor, C.J.: Using local geometry to build 3d sulcal models. In: Kuba, A., Sámal, M., Todd-Pokropek, A. (eds.) IPMI 1999. LNCS, vol. 1613, pp. 196–209. Springer, Heidelberg (1999)
22. Fleute, M., Lavallée, S.: Building a complete surface model from sparse data using statistical shape models: Application to computer assisted knee surgery. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, pp. 879–887. Springer, Heidelberg (1998)
23. Christensen, G.E., Rabbitt, R.D., Miller, M.I.: Deformable templates using large deformation kinematics. IEEE Transactions on Image Processing 5(10), 1435 (1996)
24. van de Giessen, M., Vos, F.M., Grimbergen, C.A., van Vliet, L.J., Streekstra, G.J.: An efficient and robust algorithm for parallel groupwise registration of bone surfaces. Med Image Comput Comput Assist Interv 15(3), 164–171 (2012)
25. Granger, S., Pennec, X.: Multi-scale EM-ICP: A fast and robust approach for surface registration. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 418–432. Springer, Heidelberg (2002)
26. Vroemen, J., Dobbe, J., Jonges, R., Strackee, S., Streekstra, G.: Three-dimensional assessment of bilateral symmetry of the radius and ulna for planning corrective surgeries. J. Hand Surg. Am. 5(37), 982–988 (2012)
27. Dobbe, J., du Pré, K., Kloen, P., Blankevoort, L., Streekstra, G.: Computer-assisted and patient-specific 3-d planning and evaluation of a single-cut rotational osteotomy for complex long-bone deformities. Med. and Bio. Eng. and Comp. 49(12), 1363–1370 (2011)
28. Koenderink, J.J., van Doorn, A.J.: Surface shape and curvature scales. Im. and Vis. Computing 10(8), 557–564 (1992)

An Inference Language for Imaging

Stefano Pedemonte[1,2], Ciprian Catana[1], and Koen Van Leemput[1,2,3]

[1] Athinoula A. Martinos Center for Biomedical Imaging, MGH/Harvard, MA, USA
[2] Department of Information and Computer Science, Aalto University, Finland
[3] Department of Applied Mathematics and Computer Science,
Technical University of Denmark, Denmark

Abstract. We introduce iLang, a language and software framework for probabilistic inference. The iLang framework enables the definition of directed and undirected probabilistic graphical models and the automated synthesis of high performance inference algorithms for imaging applications. The iLang framework is composed of a set of language primitives and of an inference engine based on a message-passing system that integrates cutting-edge computational tools, including proximal algorithms and high performance Hamiltonian Markov Chain Monte Carlo techniques. A set of domain-specific highly optimized GPU-accelerated primitives specializes iLang to the spatial data-structures that arise in imaging applications. We illustrate the framework through a challenging application: spatio-temporal tomographic reconstruction with compressive sensing.

1 Introduction

Probabilistic reasoning combines deductive logic with the capacity of probability theory to handle uncertainty, providing an expressive formalism with a broad range of applications in many areas of artificial intelligence and machine learning. Stochastic programming languages address the model-building process by giving a formal language which provides simple, uniform, and re-usable descriptions of a wide class of models, and supports generic inference techniques [1,2,3]. Probabilistic graphical models express explicitly the structure of probabilistic models by means of a graph, constituting a natural data structure for the design of stochastic programming languages.

The iLang framework is aimed at enabling the construction of models for imaging applications, focusing in particular on volumetric biomedical imaging. In this domain, probabilistic graphical models have been employed recently in a number of applications including image segmentation, tomographic reconstruction and multi-modal image processing. The *integrated modeling paradigm* has emerged in the work of K. Van Leemput [4], J. Ashburner [5], B. Fischl [6] and others in the context of medical image classification and alignment, adopting a model-based approach to devise algorithms for the joint estimation of multiple model parameters. Other instances of the integrated probabilistic modeling paradigm include the fusion of functional and structural information for the purpose of inferring anatomical-functional networks of the brain [7], the fusion of

M.J. Cardoso et al. (Eds.): BAMBI 2014, LNCS 8677, pp. 61–72, 2014.

information from MRI and PET [8] and the use of population-derived information for intensity-based classification of image structures [9]. In the aforementioned publications, probabilistic graphical models are employed for the purpose of describing the models and aiding the derivation of the symbolic expressions that the models imply; ad-hoc algorithms for maximum-a-posteriori inference are devised based on the resulting symbolic expressions. The iLang framework aims at enabling, under the integrated modeling paradigm, the construction of algorithms that incorporate image formation, motion correction, registration, classification, de-noising and other basic imaging tasks. The iLang framework addresses imaging as probabilistic reasoning; it includes a mechanism for the description of the model, i.e. a modeling language, a mechanism for the definition of inference queries, i.e. an inference language, and an inference engine that utilizes the data structures produced by the interpreter of the modeling language to perform inference. By using a formal modeling language, the computer gains the concept of a probabilistic model. Endowing the numerical representations of the probabilistic models with graph structures, then, enables the automated synthesis of efficient inference algorithms. The modeling language of iLang is based on language primitives designed for the construction of directed and indirected probabilistic graphical models. The inference engine of iLang addresses, without lack of generality, maximum probability and posterior sampling inference queries. The design of the language and of the data structures for the representation of the models yields a graph-based message-passing system that supports algorithms for maximum probability and posterior sampling. We describe two algorithms currently implemented in iLang: an algorithm for maximum probability estimation based on the Alternating Direction Method of Multipliers (ADMM) and an algorithm for posterior sampling of high dimensional models based on Hamiltonian Markov Chain Monte Carlo.

2 Methods

Imaging problems are not different, at the abstract level of probabilistic reasoning, from other computational problems that arise in artificial intelligence and machine learning, although imaging problems have two salient characteristics:

1. Imaging data is often very high dimensional;
2. An underlying structure, in computational problems related to imaging, arises from the spatial organization of the imaging data.

The high dimensionality may prohibit the use of certain classes of algorithms such as posterior sampling techniques. In imaging applications, the system matrices are often too large to be explicitly evaluated and stored in memory. The underlying structure, however, often can be exploited to evaluate efficiently matrix-vector multiplications on the fly and to increase the performance of the inference algorithms. A design challenge arises: abstracting imaging problems into the framework of probabilistic reasoning, therefore enabling the use of general purpose inference algorithms, while exploiting the underlying structure that arises from the spatial organization of the imaging data.

In iLang, the modeling language, the inference engine, and the inference query language are implemented as modules for the Python programming language. Language primitives constitute the units for the definition of probabilistic graphical models. As explained in the next section, the language primitives are based on a library of high performance geometric primitives which incapsulate the computations that emerge from the spatial structure of the imaging data (see Fig. 1).

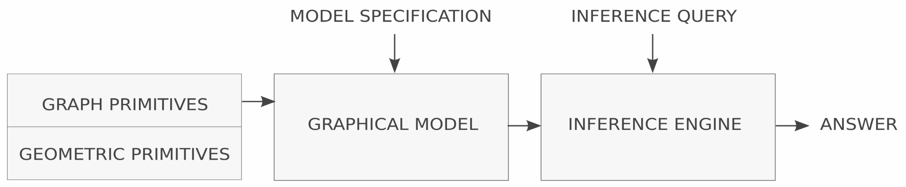


Fig. 1. The iLang probabilistic reasoning framework. The modeling language enables the definition of probabilistic graphical models using simple graph primitives. The inference engine infers the state of variables of a probabilistic graphical model. Imaging specific graph primitives are based on a library of high performance geometric primitives.

2.1 The Modeling Language

Models in iLang are constructed by defining a set of variables and specifying their interaction by means of a set of graph primitives. The following sections explain the rationale of the design of the iLang modeling language (section 2.1), describe the graph primitive construct (section 2.1), the model specification mechanism (section 2.1) and the geometric primitives that underly the graph primitives (section 2.1).

Set-Of-Rules on a Graph. A probabilistic model expresses the joint probability distribution associated to a set of variables. One question that arises when designing a software framework for probabilistic reasoning is how to define a numerical representation of a probabilistic model and whether such representation enables the construction of efficient inference algorithms. Let us consider the following approaches to the numerical representation of probabilistic models:

1. *tabulation:* A table expresses the probability of each possible state of the variables (in case of discrete variables).
2. *symbolic representation:* The joint probability distribution of the variables is expressed by a mathematical formula.
3. *set-of-rules:* A computer program returns the joint probability for a given configuration of the variables.

Tabulation is only viable for small problems and for discrete variables. Symbolic representation of probabilistic models is an active research topic [10], enabling low memory representations of models and the automated computation of the derivatives of the probability functions via symbolic expression manipulation. While the symbolic approaches yield efficient and flexible tools for probabilistic reasoning [3], their use is currently limited to small problems where the system matrices can be explicitly evaluated. The third approach consists in writing a computer program that evaluates the probability associated to a given configuration of the variables and eventually the log-conditional probability functions associated to subsets of the variables and the derivatives of the log-probability functions. The most common approach to probabilistic modeling in medical imaging is based on the *set-of-rules* representation. One describes the model in symbolic form with pen-and-paper and manipulates the symbolic expressions to obtain expressions of the required conditionals, marginals and derivatives. A computer program that evaluates such expressions is then crafted. The conditional independencies of a probabilistic model can be represented by means of a graph (i.e. a probabilistic graphical model). The set of conditional independencies corresponds to the factorization of the joint probability distribution associated to the variables of the model. The explicit representation of the set of conditional independencies via a graph provides insight of the probabilistic model. The graph is often utilized, therefore, in order to aid the pen-and-paper symbolic expression manipulation and crafting of the computer programs: through the properties of the graph, one can tell which variables (Markov blanket) and factors contribute to the conditional probability distribution of a subset of the variables. The iLang framework adopts the model representation approach 3, in conjunction with a data structure based on the graph of the probabilistic model. Informing the computer software of the conditional independence structure of the model introduces many advantages. The combination of the set-of-rules approach and the graph, while allowing maximum flexibility, simplifies the model specification process, provides a mechanism for code encapsulation and provides a data structure suitable for the automated synthesis of inference algorithms. The core data-structure representing an iLang model is a graph, defined by a set of graph primitives. A graph primitive defines the interaction between a set of variables in terms of sets of rules for the computation of log-conditional probabilities and their derivatives, as described in the next section. Pen-and-paper symbolic manipulation is still part of the model definition process, however occurring only at the stage of designing a graph primitive.

Graph Primitives. A graph primitive of the iLang modeling language expresses the interaction between a set of variables. Variables associated to a graph primitive are objects with a *name* property and a *value* property. The graph primitive object exposes, for each of the internal variables, one to four methods that return 1) the log conditional probability; 2) the gradient of the log conditional probability; 3) the Hessian of the log conditional probability; 4) the proximity map of the log conditional probability. This is depicted in Fig. 2. A graph primitive is defined by subclassing a base object of type *GraphPrimitive*; defining a dictionary with the names of the variables; a dictionary that specifies the directed or

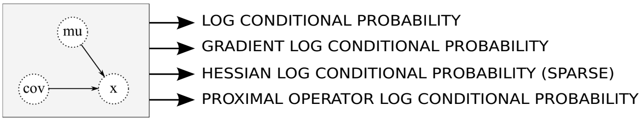

mu
cov
x
LOG CONDITIONAL PROBABILITY
GRADIENT LOG CONDITIONAL PROBABILITY
HESSIAN LOG CONDITIONAL PROBABILITY (SPARSE)
PROXIMAL OPERATOR LOG CONDITIONAL PROBABILITY

Fig. 2. Interface of a graph primitive: a graph primitive encodes the dependence amongst variables by specifying methods to compute the log conditional probabilities of each variable. Optionally, the graph primitive exposes methods to compute the first and second derivatives and the proximal operator of the log conditional probabilitties.

indirected graph structure; and by implementing interface methods according to a simple predefined naming convention. The example that follows specifies a graph primitive that encodes a multivariate Gaussian probability distribution $p(x|mu, cov) = \mathcal{N}(x; mu, cov)$:

```
class MultivariateGaussian(GraphPrimitive):
    variables  = {’x’:’continuous’,’mu’:’continuous’,’cov’:’continuous’}
    dependencies = [[’mu’,’x’,’directed’],[’cov’,’x’,’directed’]]
    preferred_samplers = {’x’:[’HamiltonianMCMC’]}

    # graph primitive interface
    def log_conditional_probability_x(self,x):
        hessian = self._compute_hessian()
        mu = self.get_value(’mu’)
        return -.5*numpy.dot(numpy.dot((x-mu),hessian),(x-mu).T)

    def log_conditional_probability_gradient_x(self,x):
        hessian = self._compute_hessian()
        mu = self.get_value(’mu’)
        return -.5*numpy.dot((x-mu),hessian+hessian.T)

    def log_conditional_probability_hessian_x(self,x):
        hessian = self._compute_hessian()
        return hessian

    # utility:
    def _compute_hessian(self):
        cov     = self.get_value(’cov’)
        self._hessian = numpy.linalg.inv(cov)
        return self._hessian
```

Note, in the example, that variables are defined as discrete or continuous. Further classes of the iLang variables will be added in future implementations, such as symmetric, positive definite and chordal matrices. The graph primitives currently implemented in iLang are reported in Fig. 3.

Model Specification. A model is specified by instantiating an object of type *GraphicalModel* and naming the variables of the model. The dependence between the variables is specified by connecting the variables by means of graph primitives, as in the example that follows.

```
graph = ilang.GraphicalModel()
graph.add_variables([’var1’,’var2’,’var3’])
graph.add_model(MultivariateGaussian,{’var1’:’x’,’var2’:’mu’,’var3’:’cov’})
graph.set_given({’var2’:numpy.zeros([1,5]),’var3’:numpy.eye(5)})
```

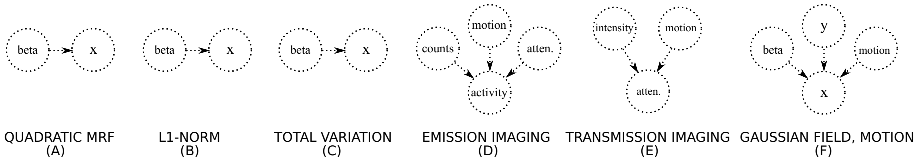


Fig. 3. Graph primitives currently implemented in iLang

In this example, the object *graph* represents a probabilistic graphical model with 3 variables: *var1*, *var2*, *var3*; *var2* and *var3* have given values and *var1* is a 5-dimensional random variable with probability distribution $p(var1|var2, var3) = \mathcal{N}\,(var1; var2, var3)$. Note that the correspondence between the variables of the graph and the inner variables of the graph primitive has been specified in the *add_model* function call. The *graph* object exposes the methods that are required to perform inference; the internal machinery of the *graph* object translates the names of the variables, calling the methods of the graph primitives as required.

Geometric Primitives. The graph primitives for imaging make use, internally, of efficient GPU-accelerated routines that perform common image processing tasks. Currently:

- Rigid spatial transformations
- Ray-tracing
- Image re-sampling
- FFT-IFFT
- Finite difference operator

Such geometric primitives enable a wide range of models and algorithms. The experiments section highlights how the spatial transformation, resampling and finite difference geometric primitives come into play to define a graph primitive that enables spatio-temporal tomography.

2.2 The Inference Engine

The probabilistic graph object provides all the methods required to perform inference. These include (proxy) methods to compute the log conditional probability of each of the variables and their first and second derivatives; methods to compute properties of the graph, such as the global Markov properties; and methods for the manipulation of the graph. Currently, the iLang framework implements an algorithm based on the Alternating Direction Method of Multipliers (ADMM) for maximum probability inference and an algorithm based on Hamiltonian Markov Chain Monte Carlo for posterior sampling.

Maximum Probability. The algorithm for maximum probability estimation is based on a combination of the Iterated Conditional Modes (ICM) [11] algorithm and the ADMM algorithm [12]. The value of the variables of the model that maximizes the joint probability is computed by maximizing, in turn, the conditional probability distribution of each of the variables (ICM). The optimization of each conditional probability is performed by means of the ADMM algorithm. ADMM, developed in the context of convex optimization [12], has the advantage of enabling the use of non-differentiable factors, such as the models (B) and (C) in Fig. 3, using the proximity operators of the factors in place of the first derivatives. In order to apply ADMM to optimize the log conditional probability for each of the variables, the inference engine performs a transformation of the graph, consisting in extracting the Markov blanket of the variable and transforming it into a Forney-style augmented factor graph [13], as exemplified in Fig. 4 for variable x. The ADMM algorithm then consists in a message passing algorithm over the factor graph. The Forney-style factor graph represents the augmented Lagrangian of the local optimization problem expressed in consensus form [12,13]. The Forney-style factor graph is a bipartite graph obtained by placing on the right side one node for each factor of the cost function (3 nodes in the example, corresponding to 3 graph primitives) and one node on the left side, encoding an equality constraint. The factor graph in this form expresses the augmentation of the optimization problem with variables x_1, x_2, x_3 (see Fig. 4-right). The point of the augmentation is that edge variables attached to the same equality constraint must ultimately equal each other, but they can temporarily be unequal while they separately try to satisfy different cost functions on the left. Finally, the problem is augmented with one variable for each edge connecting the two sides of the bipartite factor graph: the Lagrangian multipliers y_1, y_2, y_3. The ADMM algorithm consists in the exchange of the following messages (see [12] for the derivation of the messages):

$$x_k^{n+1} := \arg\min_v f_k(v) + \frac{\rho}{2}\|v - y_k^n + x^n\|_2^2 \tag{1}$$

$$x^{n+1} := \frac{1}{N}\sum_{k=1}^{N} x_k^{n+1} + \frac{1}{\rho N}\sum_{k=1}^{N} y_k^n \tag{2}$$

$$y_k^{n+1} := y_k^n + x_k^{n+1} - x^{n+1}, \tag{3}$$

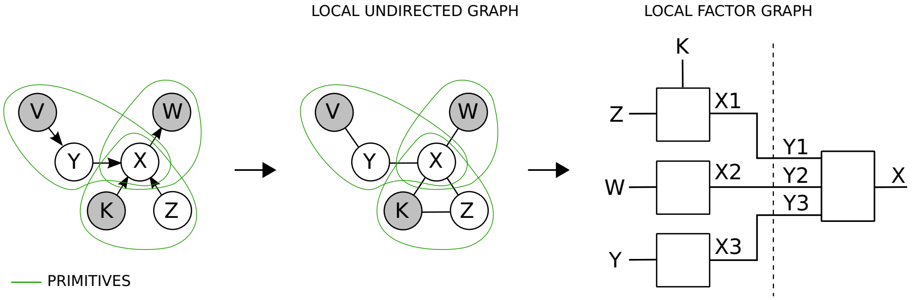


Fig. 4. Transformations of a probabilistic graphical model. Left: a directed probabilistic graphical model; center: moralized undirected gaph; right: Forney-style factor graph utilized by the inference engine of iLang.

where f_k is the k-th factor (with $k = \{1, 2, 3\}$ in the example) and ρ is the augmented Lagrangian regularization parameter (the default value is $\rho = 0.1$, see [12] for a discussion on the selection and adaptation of ρ). The splitting introduced by data augmentation enables the use of non-smooth factors. If f_k is smooth (the graph primitive corresponding to factor k exposes a method to compute the gradient of the log conditional probability), the inference engine performs the minimization using, by default, the L-BFGS Quasi-Newton algorithm, or the Newton algorithm if the graph primitive exposes a method to compute the Hessian of the log conditional probability. If the factor is non-smooth, the inference engine sets x_k^{n+1} by evaluating the proximity operator of f_k, calling the proximity operator method of the underlying graph primitive.

Posterior Sampling. The posterior sampling algorithm currently implemented in iLang is based on Markov Chain Monte Carlo. Each of the variables of the graph are sampled in turn by sampling from their conditional probability distributions (Gibbs sampling). The samples from each of the conditional probability distributions are obtained by means of various MCMC techniques, depending on the methods exposed by the factors of each conditional probability distribution. A local factor graph analogous to Fig. 4-right is constructed; if all the graph primitives connected to variable x expose methods to compute the gradient of the conditional probability of x, the MCMC algorithm uses Hamiltonian dynamics [14] with gradient equal to the sum of the gradients returned by each primitive. The local set of variables is augmented with momentum variable q and each new sample of x is obtained by sampling a candidate of q from a normal probability distribution and then by sampling x conditionally to q as follows (see [14]):

$$q^{n+1}|x^n \approx p(q^{n+1}|x^n) = p(q^{n+1}) = \mathcal{N}\left(q^{n+1}|0, M\right) \tag{4}$$

$$x^{n+1}|q^{n+1} \approx p(x^{n+1}|q^{n+1}), \tag{5}$$

samples of x^{n+1} from $p(x^{n+1}|q^{n+1})$ are obtained by integrating the Hamiltonian dynamics over fictitious time τ from the initial values q^{n+1} and x^n. The integration is performed using the leapfrog method:

$$q(\tau + \frac{\epsilon}{2}) = q(\tau) + \frac{\epsilon}{2}\nabla_x f\left(x(\tau)\right) \tag{6}$$

$$x(\tau + \epsilon) = x(\tau) + \epsilon M q\left(\tau + \frac{\epsilon}{2}\right) \tag{7}$$

$$q(\tau + \epsilon) = q(\tau + \frac{\epsilon}{2}) + \frac{\epsilon}{2}\nabla_x f\left(x(\tau + \epsilon)\right), \tag{8}$$

with a certain number of steps, with step size ϵ, to give proposed moves x^* and q^* and accepting or rejecting according to the Metropolis Hastings criterion, as specified in [14]. Here ∇_x denotes the gradient of factor f and M is a weight matrix. The weight matrix, by default, it set to the identity, unless all the factors expose methods to compute the Hessian, in which case it is set to the sum of the

Hessian terms, producing a piecewise constant Riemannian Manifold Hamiltonian MCMC algorithm [14]. Although the choice of the parameter ϵ is in general critical in order to obtain high acceptance ratios, especially in high dimensions, setting M to the Hessian of the factor f, as discussed in [14], enabling the algorithm to scale to high dimensions and relaxes the choice of ϵ. This is the default mode of iLang if all the factors expose the Hessian method. The default value of ϵ is 0.1.

3 Motion-aware Positron Emission Tomography

In PET imaging, the low number of photon counts per unit time imposes long acquisition times (several minutes). During the acquisition, the subject moves, determining blurring and ghosting effects in the reconstructed images. Although attempts have been made to measure the motion of the subject during the acquisition of the PET data by using motion detection devices, the problem is still largely unsolved. The problem can be formulated in the probabilistic framework as follows, in the case, applicable to brain imaging, of rigid motion. Although the activity in the imaging volume changes over time due to motion, let us assume, disregarding pharmacokinetics in this first instance of spatio-temporal model, that the rate of emission in the frame of reference that moves rigidly with the head of the patient is constant. Assuming that the only source of uncertainty associated to the measurements is the inherent uncertainty due to photon counting, the conditional probability distribution associated to the photon counts at time t, given the motion parameters at time t and the activity in the reference frame, is a Poisson distribution. Let us denote with $q_d^{[t]}$ the photon counts along line of response (LOR) d at time t; with $z^{[t]} = \{z_1^{[t]}, z_2^{[t]}, \dots, z_d^{[t]}\}$ the vector of the photon counts at time t; with $A = \{a_{bd}\}$ the matrix of the probabilities that an event emitted in voxel b is detected in LOR d; with $R_{\gamma^{[t]}}$ the rigid transformation at time t, parameterized by parameters $\gamma^{[t]}$ and with $\mathcal{P}$ the Poisson distribution:

$$p(z^{[t]}|\lambda, R_{\gamma^{[t]}}) = \prod_d \mathcal{P}(\sum_b a_{bd}[R_{\gamma^{[t]}}\lambda]_b, z_d^{[t]}) \tag{9}$$

Let us assume a sparsifying total-variation prior probability distribution for the activity:

$$p(\lambda|\beta) \propto e^{-\beta\|\nabla\lambda\|_1} \tag{10}$$

Denoting by $\bar{1}$ the vector of 1's, the gradient of eq. (9) is given by (see [8]):

$$\frac{\partial}{\partial\lambda_b} \log p(\lambda|z_{[t]}, R_{\gamma^{[t]}}) = -\sum_t R^T_{\gamma^{[t]}} A^T \bar{1} + \sum_t R^T_{\gamma^{[t]}} A^T \frac{z^{[t]}}{AR_{\gamma^{[t]}}\lambda} \tag{11}$$

Let us assume that the motion parameters $\gamma^{[t]}$ are unpredictable, i.e. that the motion parameter $\gamma^{[t]}$ is a priori independent from the motion parameter $\gamma^{[t']}$,

$t \neq t'$. By the chain rule of differentiation, the derivative of the log conditional probability of the i-th motion parameter at time t is given by:

$$\frac{\partial \log p(\gamma^{[t]}|z^{[t]},\lambda)}{\partial \gamma_i^{[t]}} = \sum_d - \left[A\left[\frac{\partial R^T_{\gamma^{[t]}}\lambda}{\partial\gamma_i^{[t]}}\right]\right]_d + z_d^{[t]} \frac{\left[A\left[\frac{\partial R^T_{\gamma^{[t]}}\lambda}{\partial\gamma_i^{[t]}}\right]\right]_d}{\left[AR^T_{\gamma^{[t]}}\lambda\right]_d} \tag{12}$$

Optimization of the joint probability with respect to the model parameters is not trivial due to the non-differentiability of the prior and to the non-negativity constraint (here not expressed explicitly) of λ. In iLang, the calculations of eq. (11) and (12) are encapsulated in the graph primitive (D) of Fig. 3 and the graph primitive (C) of Fig. 3 implements the proximity operator for the total variation prior (i.e. soft thresholding of the image gradient - see [12]). The model is encoded in iLang as follows:

```
graph = ilang.GraphicalModel()
graph.add_variables(’lambda’)
for t in range(Nt):
    graph.add_variable(’z’+str(t))
    graph.add_variables(’gamma’+str(t))
    graph.add_model( ilang.Primitives.PET_rigid_motion , ..
      {’lambda’:’activity’,’gamma’+str(t):’motion’,’z’+str(t):’counts’})
    graph.set_given(’z’+str(t))
    graph.set_value(’z’+str(t), sinograms[t])
graph.add_variable{’beta’}
graph.add_model(ilang.Primitives.TotalVariation , ..
  {’lambda’:’x’,’beta’:’sparsity’})
graph.set_given(’beta’)
graph.set_value(’beta’,0.1)
```

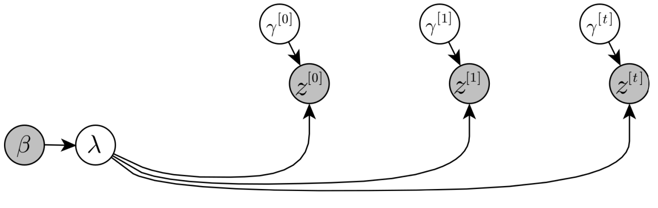


Fig. 5. Graph generated by iLang for motion-aware Positron Emission Tomography

where sinogram is a list of N_t sinogram arrays. This produces the graph of Fig. 5. Inference is performed as follows:

```
sampler = ilang.Sampler(graph)
sampler.maximum_probability(max_iterations=100)
activity_estimate = sampler.get_last_sample(’lambda’)
```

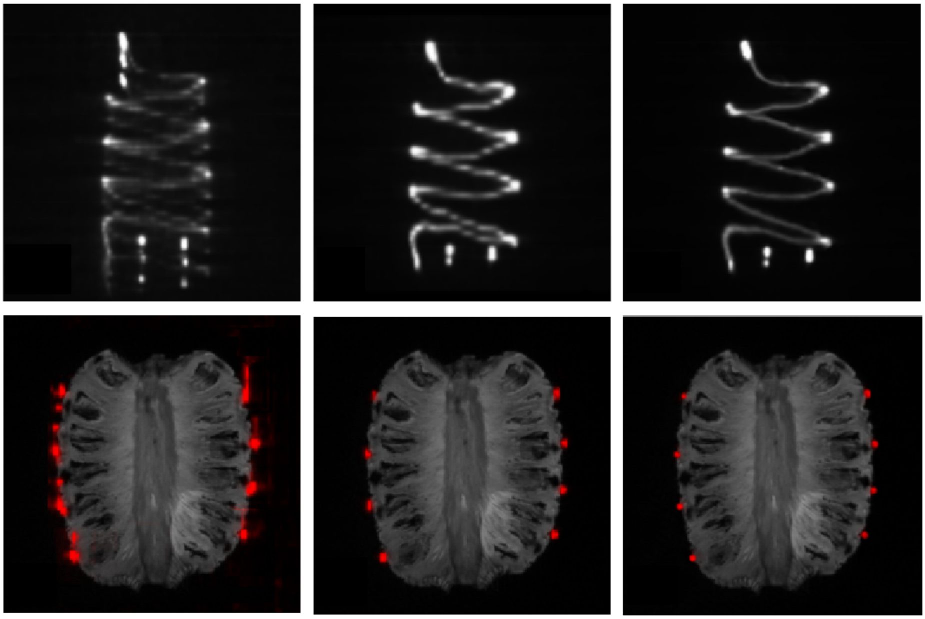


Fig. 6. Motion-Aware PET: Reconstructions obtained by imaging an FDG-filled capillary source wrapped around a pineapple. From left to right: no motion correction; motion estimation; motion estimation and total-varation prior. Top row: volume rendering; bottom row: representative slice - overlay of PET and MR.

4 Conclusion

The iLang software framework enables probabilistic reasoning in volumetric imaging, simplifying the definition of complex imaging models. Endowing the numerical representation of probabilistic models with a graph enables the automated synthesis of efficient inference algorithms. The inference engine of iLang enables the definition of non-smooth constraints such as non-negativity and sparsity. The iLang software constitutes a unified framework for multi-modal imaging that enables the integration of image formation, registration, de-noising and other image processing tasks. The application reported in section 3 constitutes a novel powerful imaging paradigm, where motion is considered a nuisance variable and estimated from the PET emission data under the assumption of sparsity.

5 Download

The iLang software is distributed with a permissive open source license at `http://ilang.github.io`

Acknowledgements. This research was supported by the NIH NCRR (P41-RR14075), the NIH NIBIB (R01EB013565), and TEKES (ComBrain).

The authors would like to thank Paulina Golland and the MIT EECS/CSAIL journal club for the useful introduction to the ADMM algorithm.

References

1. Goodman, N., et al.: Church: a language for generative models. arXiv 1206.3255 (2012)
2. Stan modeling language users guide and reference manual. Technical report (2014)
3. Patil, A., Huard, D., Fonnesbeck, C.: Pymc: Bayesian stochastic modelling in python. J. Stat. Softw. 35(4), 1–81 (2010)
4. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based tissue classification of mr images of the brain. IEEE Trans. Med. Imaging 18(10), 897–908 (1999)
5. Ashburner, J., Andersson, J.L., Friston, K.J.: Unified segmentation. Neuroimage 26(3), 839–851 (2005)
6. Fischl, B., Salat, D.H., van der Kouwe, A.J., Makris, N., Segonne, F., Quinn, B.T., Dale, A.M.: Sequence-independent segmentation of magnetic resonance images. Neuroimage 23, 69–84 (2004)
7. Venkataraman, A. Rathi, Y.K.M.W.C.G.P: Joint modeling of anatomical and functional connectivity for population studies. IEEE Trans Med Imaging 31(2), 164–82 (2012)
8. Pedemonte, S. Bousse, A.H.B.A.S.O.S.: 4-d generative model for pet/mri reconstruction. MICCAI 2011 14(1), 581–588 (2011)
9. Menze, B.H., van Leemput, K., Lashkari, D., Weber, M.-A., Ayache, N., Golland, P.: A generative model for brain tumor segmentation in multi-modal images. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part II. LNCS, vol. 6362, pp. 151–159. Springer, Heidelberg (2010)
10. Bergstra, J. Breuleux, O.B.F.L.P.P.R.D.G.T.J.W.F.D., Bengio, Y.: Theano: A cpu and gpu math expression compiler. Proceedings of the Python for Scientific Computing Conference (SciPy) 48(3), 259–302 (2010)
11. Besag, J.: On the statistical analysis of dirty pictures. J. of the Royal Stat. Soc. Series B (Methodological) 48(3), 259–302 (1986)
12. Boyd, S. Parikh, N.C.E.P.B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning 3(1), 1–122 (2010)
13. Forney Jr, G.: Codes on graphs: Normal realizations. IEEE Transactions on Information Theory 47(2), 520–548 (2001)
14. Girolami, M., Calderhead, B.: Riemann manifold langevin and hamiltonian monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(2), 123–214 (2014)

An MRF-Based Discrete Optimization Framework for Combined DCE-MRI Motion Correction and Pharmacokinetic Parameter Estimation

Monica Enescu[1], Mattias P. Heinrich[3], Esme Hill[2], Ricky Sharma[2], Michael A. Chappell[1], and Julia A. Schnabel[1]

[1] Institute of Biomedical Engineering, University of Oxford, UK
[2] Department of Oncology, University of Oxford, UK
[3] Institute of Medical Informatics, University of Lüebeck, Germany

Abstract. Dynamic contrast-enhanced MRI (DCE-MRI) images are increasingly used for assessing cancer treatment outcome. These time sequences are typically affected by motion, which causes significant errors in tracer kinetic model analysis. Current intra-sequence registration methods for contrast enhanced data either assume restricted transformations (e.g. translation) or employ continuous optimization, which is prone to local optima. In this work, we propose a new approach to DCE-MRI intra-sequence registration and pharmacokinetic modelling, which is formulated in an MRF optimization framework. The complete $4D$ graph corresponding to a DCE-MRI sequence is reduced to a concatenation of minimum spanning trees, which can be optimized more efficiently. To address the changes due to contrast, a data cost function which incorporates pharmacokinetic modelling information is formulated. The advantages of this method are demonstrated on 8 DCE-MRI image sequences of patients with advanced rectal tumours, presenting mild to severe motion.

1 Introduction

DCE-MRI has become an important tool for assessing early phase clinical trials of cancer therapy, as it can measure in vivo tumour vasculature changes that occur due to treatment. The underlying tissue physiology is typically derived from the DCE-MRI image signal by fitting a pharmacokinetic (PK) model to the contrast enhancement-time curve on a voxel-by-voxel basis. Typically, tissue perfusion, permeability and the volume occupied by tumour cells are obtained in terms of PK model parameters. As a DCE-MRI acquisition takes several minutes, with volumes being acquired every $5 - 10$ seconds, the resulting time sequence is inherently affected by patient and physiological motion. This motion may introduce significant errors to the per-voxel PK model fitting, as anatomical features of interest might move to different voxel locations in subsequent volumes. To correct for this motion, image registration is needed. DCE-MRI registration is a particularly challenging problem, as observed changes throughout the time series can be either due to motion or due to contrast enhancement.

M.J. Cardoso et al. (Eds.): BAMBI 2014, LNCS 8677, pp. 73–84, 2014.

Moreover, contrast arrival can give rise to image features that were not present in the baseline image. In the literature, approaches for time series motion correction broadly fall into two categories: 1. strategies that try to alleviate the effects of contrast enhancement i.e. by using a multi-modal similarity metric [1], or by restricting the applied transformation [2]; 2. strategies that implicitly derive [3] or explicitly assume [4,5,6] a model of contrast enhancement which is used in the registration algorithm. Among the most prominent approaches, we note the work of Buonaccorsi et al. [5], who are among the first to explore an explicit kinetic model-based registration. In that work, PK parameter estimation and registration to the model predicted sequence are performed iteratively. Their work is mainly limited by allowing only 3D translation transforms. In a more recent approach, Bhushan et al. [6] address this issue by proposing a simultaneous non-rigid motion correction and PK parameter estimation method. However, as their approach uses a Gauss-Newton optimization, this method is sensitive to initialization and is likely to be trapped in local optima. A data-driven approach, presented by Melbourne et al. [3], proposes modelling the time series data using principal component analysis. The underlying assumption is that the first few principal components will contain information about contrast enhancement trends and the remaining principal components contain noise related to motion. This assumption is valid for small peristaltic motion, but does not hold in the case of larger and periodic motion (e.g. breathing).

In this work, we propose a new framework for combined DCE-MRI intra-sequence registration using discrete optimization and PK parameter estimation (DireP). The problem is formulated using a Markov random field (MRF) which involves the optimization on a $4D$ graph for each DCE-MRI sequence. This method addresses the sensitivity to initialization of continuous approaches and is less prone to local optima by offering increased flexibility over the space of possible displacements $\mathcal{L}$. To reduce the computational costs which are typically associated with discrete optimization of a full 4D graph, the nodes are connected as follows: The minimum spanning tree (MST) which best replicates the underlying anatomy of the pre-contrast image is calculated [7]. This structure is assumed to be identical in all the subsequent volumes, as they have the same anatomy as the baseline. These structures are connected through time at every node, to preserve the temporal continuity of each voxel (Fig. 1).

The paper is structured as follows. Section 2 describes the methodological contributions of this paper: First, a DCE-MRI tailored similarity metric which incorporates PK modelling information is formulated. Next, in order to reduce computational costs, we construct a reduced $4D$ graph corresponding to the DCE-MRI sequence, and optimize it using a message passing approach. To our knowledge, this is the first work to perform combined DCE-MRI motion correction and PK analysis, where the registration is performed in a discrete optimization framework. Section 3 describes the results of the proposed algorithm (DireP), on both synthetic and real DCE-MRI images of rectal cancer. DireP is compared to a recent DCE-MRI non-rigid registration algorithm (referred to

Method1) [6] which employs continuous optimization. We conclude this paper with Section 4, and present future work plans in Section 5.

2 Methods

Discrete optimization is typically formulated as an MRF labelling problem [8]. For deformable image registration purposes, a graph is defined in which the nodes $p \in \Omega$ represent voxels or groups of voxels, and the edges connect voxels with similar anatomical features and spatial proximity. For every node p, there is a set of labels $l_p \in \mathcal{L}$ which represent possible discrete displacements of the source image volume with respect to the target image volume. Finding the optimum displacement at each voxel equates to finding the labelling that minimizes the MRF energy function:

$$E(l) = \sum_{p \in \Omega} C_D(l_p) + \gamma \sum_{(p,q) \in \mathcal{N}} C_R(l_p, l_q) \tag{1}$$

The unary term represents the data cost C_D, which measures the similarity of a voxel in the target image to the corresponding voxel in the source image displaced with l_p. The pairwise term represents the regularization cost $C_R(l_p, l_q)$ and is used to smooth the displacements of directly connected voxels $(p, q) \in \mathcal{N}$. Here, $\mathcal{N}$ represents the neighborhood of a voxel, as given by the MST. γ weights the amount of regularization.

Methods to solve the labelling problem can be roughly divided into two categories: graph-cuts and message passing approaches. Popular graph-cuts algorithms include α-expansion [9]. Depending on the complexity of the graph to be optimized, message passing can range from dynamic programming, over loopy belief propagation (LBP) [10], to tree-reweighted message passing (TRW-S) [11]. For an overview of discrete methods for deformable registration, we refer the reader to the work of Sotiras et al. [12].

In this work, each image volume of the DCE-MRI sequence is registered to the pre-contrast volume by optimizing a reduced $4D$ graph corresponding to the underlying anatomy. To address the intensity differences caused by contrast inflow, a DCE-MRI tailored data cost function incorporating PK information is proposed. The optimization is performed using belief propagation. An overview of the entire algorithm for finding the optimal displacement and updating the PK model can be found in Algorithm 1.

2.1 Data Cost Calculation Using Pharmacokinetic Model Prediction

As mentioned above, the DCE-MRI intensities of different volumes in the image sequence are not comparable using standard similarity metrics such as the sum of squared differences (SSD). To address this issue, we propose to compare the similarity between the intensities at the volume to be registered I_{t_i}, and the PK model predicted intensity at that volume $PK(I_{t_0}, K^{trans}, v_e, t_i)$:

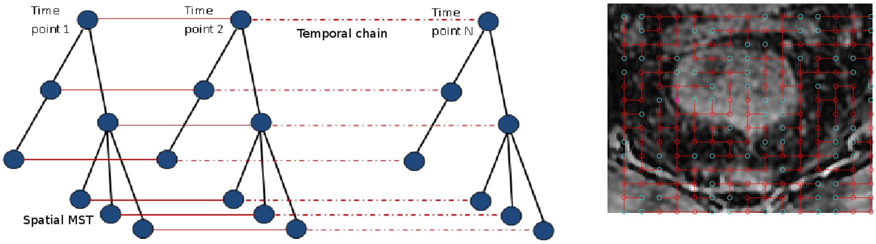
$$C_D(l_p) = SSD(I_{t_i}(p + l_p), PK(I_{t_0}(p), K^{trans}(p), v_e(p), t_i)) \tag{2}$$

Here, t_i is the current time point of the DCE-MRI sequence, t_0 is the first time point, and K^{trans} and v_e are parameters of the PK model. To predict the appearance of the baseline I_{t_0} at t_i, an initial estimate of the PK parameters K^{trans}, v_e is required. This initial estimate is obtained by least squares (lsq) fitting of the model to the data.

PK parameter estimation was performed using the standard Tofts model [13] with the Orton [14] bi-exponential arterial input function with population averaged parameters. The Tofts model offers an estimate of tissue perfusion through K^{trans}, the transfer constant between blood plasma and the extravascular extracellular space (EES), which is defined as the product between k_{ep}, the rate constant between the EES and plasma, and v_e, the fractional volume of the EES.

2.2 Optimization on the Reduced 4D Graph

For the DCE-MRI motion correction problem, the graph to be optimized can be reduced to temporal concatenation of identical spatial trees (Fig. 1). The pre-contrast image anatomy is taken as a reference, against which all the subsequent volumes are aligned. We assume that they obey the same anatomy and differ only by contrast enhancement or motion. Under this assumption, we estimate the unique MST of the pre-contrast volume using Prim's algorithm [7] akin to the work of Heinrich et al. [15]. This structure is replicated for each subsequent volume, and the temporally corresponding nodes are additionally connected across time, as they represent the same anatomical landmark over time. For each individual tree, the optimization is possible using belief propagation on the MST [16]. At each node p, a message vector m_p containing the cost of the best displacement l_p can be found given the displacement of its parent q, l_q, and the messages from its children c, m_c:



(a) (b)

Fig. 1. (a) An illustration of the graphical structure used in this algorithm. While the spatial connectivity is captured by minimum spanning trees (MST), the temporal continuity is captured by temporal chains. (b) An example of MST for a 2D slice through the mesorectal area.

$$m_p^{tree}(l_q) = \min_{l_p}(C_D(l_p) + \gamma_{sp} C_R(l_p, l_q) + \sum_c m_c(l_p)) \tag{3}$$

For leaf nodes, Eq. 3 can be evaluated directly, as there are no incoming messages from children. The messages are then forward-passed from the leaves to the root node, and then backward-passed, from the root towards the leaves. For each individual temporal chain, the optimization can be performed using belief propagation on the chain [17]. The message of the current node p, m_p, can be found given the displacement of the previous node q, l_q, and the message coming from the subsequent node r, $m_r(l_p)$:

$$m_p^{chain}(l_q) = \min_{l_p}(C_D(l_p) + \gamma_{temp} C_R(l_p, l_q) + m_r(l_p)) \tag{4}$$

To optimize the reduced 4D structure, for each node we perform independent optimization on each temporal chain and on each spatial tree, and average the resulting marginals. This procedure is repeated for a number of iterations akin LBP [10], where the messages towards a node in the previous iteration are added to the marginal of that node in the current iteration. The best displacement can then be found by calculating argmin based on the marginal distribution. Although this approach is not guaranteed to converge to a global optimum, it is physically motivated, and provides a good trade-off between optimality and efficiency. The averaging of marginals from multiple graph models has previously been used in stereo processing using an approach called semi-global matching [18], yielding excellent results. The entire DireP algorithm is presented in Algorithm 1. In each optimization step (Algorithm 1, 2.2, 2.3), the full distribution of pseudo-marginals for the space of displacements is estimated.

Naively calculating the pair-wise regularization cost in Eqs. 3, 4 would require $|\mathcal{L}|^2$ computations for every voxel. To reduce this cost, we employed the min-convolution technique [10], which reduces the complexity to $|\mathcal{L}|$. To further reduce computational costs, a multi-resolution approach was employed. For each resolution level, the image is divided into non-overlapping cubic groups of voxels that are represented by a single node in the graph. The regularization term is calculated only for each group of voxels. At finer levels, the previous solution is upsampled and used as a prior for the optimization algorithm. The final (dense) solution is also obtained by upsampling using a first order spline interpolation.

Additionally, if we treat the deformation field as a velocity field, it can be transformed into a diffeomorphic mapping [15] by using the scaling and squaring method [19]. Diffeomorphism is a desired property for the deformation field, as it prevents transformations that are not physically feasible, i.e. folding. This is particularly important for soft tissue images, as it agrees with the tissue incompressibility assumption.

3 Results

3.1 Algorithm Evaluation on Synthetic Data

The discrete registration and pharmacokinetic estimation (DireP) algorithm proposed in this paper was first tested on synthetic data, where the ground truth intra-sequence motion and PK parameters are known. To simulate a realistic

Algorithm 1. DireP: Discrete motion correction and pharmacokinetic estimation

```
1. PK parameter estimation on uncorrected time series (lsq fitting)
while $n_{iterPK}$ do
  2. Groupwise registration of all volumes to the pre-contrast volume:
  2.1 *Initialize marginals and messages:*
  **foreach** *node* **do** marginal_c[ node ]=$C_D(l_p)$; marginal_t[ node ]=$C_D(l_p)$;
  message[ node ]=0;
  while $n_{iterMRF}$ do
    2.1 *Re-initialize marginals with values from previous iteration*
    2.2 **foreach timepoint** *pass messages on the corresponding spatial MST*
    *Forward pass*
    for node=leaves to root-1 do
      cost = marginal_t[node];
      message[ node ]=min-sum(cost); (see Eq. 3)
      marginal_t[ parent ] = marginal_t[ parent ]+message[ node ];
    end for
    *Backward pass*
    for node=root-1 to leaves do
      cost=marginal_t[ parent ] - message[ node ]+ message[ parent ];
      message[ node ]=min-sum(cost);
    end for
    foreach node marginal_t[ node ]=marginal_t[ node ]+message[ node ];
    2.3 **foreach node** *of the spatial MST, pass messages on the temporal chain*
    *Forward pass*
    for t=tdim-1 to 1 do
      cost = marginal_c[ $\mathrm{node}_t$ ];
      message[ $\mathrm{node}_t$ ]=min-sum(cost); (see Eq. 4)
      marginal_c[ $\mathrm{node}_{t-1}$ ] = marginal_c[ $\mathrm{node}_{t-1}$ ]+message[ $\mathrm{node}_t$ ];
    end for
    *Backward pass*
    for t=1 to tdim-1 do
      cost=marginal_c[$\mathrm{node}_{t-1}$ ] - message[ $\mathrm{node}_t$ ]+ message[ $\mathrm{node}_{t-1}$ ];
      message[ node ]=min-sum(cost);
    end for
    foreach node marginal_c[ node ]=marginal_c[ node ]+message[ node ];

    2.4 *Average the two marginals and use the result in the next iteration*
  end while
  3. Re-estimate PK parameters on corrected time series (lsq fitting),
  use in the next iteration
end while
```

dataset, synthetic images were generated as follows: 8 real DCE-MRI sequences were selected, and PK model fitting was performed on each of them. Using the resulting parameter maps and the pre-contrast image volumes, 8 model predicted sequences of size $120 \times 120 \times 52 \times 29$ each were generated. These images constituted the ground truth motion free dataset. To simulate motion, a random displacement field of size $6 \times 6 \times 4$ was generated for each image volume independently. This displacement was upsampled to the image volume size, and smoothed with a Gaussian kernel. We also applied Gaussian smoothing on the temporal dimension, as the motion of each voxel is expected to have some degree of temporal smoothness due to periodic motion patterns i.e. breathing. The parameters for the synthetic motion were chosen to obtain a diffeomorphic deformation field in the interval ±7mm along each direction. The experiments were run with $n_{iterPK} = 2$, $n_{iterMRF} = 5$ (see Algorithm 1), $\gamma_{sp} = 0.01$ and $\gamma_{temp} = 0.1$, which were empirically chosen. For both the synthetic and the real data, the label space was chosen as $L = \{0, \pm q, \pm 2q, .., \pm \frac{u}{2} q\}^3$. 3 resolution levels were used in the registration. Depending on the resolution level, u was $8, 6, 4$, with a quantization q of $2, 1, 0.5$mm. We used groups of voxels of sizes 8^3, 6^3 and 4^3, with corresponding label spaces of size 9^3, 7^3 and 5^3.

Table 1. Registration results on synthetic data. The average target registration error (TRE) and residual fitting errors in PK parameters are reported, together with the corresponding standard deviations.

Measure	Before	Method1	DireP $3D$	DireP $4D$
TRE (mm)	1.40 ± 0.72	1.02 ± 0.65	$\mathbf{0.59 \pm 0.38}$	0.62 ± 0.39
Error in k_{ep}	0.39 ± 0.03	0.15 ± 0.02	$\mathbf{0.13 \pm 0.01}$	$\mathbf{0.13 \pm 0.01}$
Error in v_e	0.20 ± 0.01	0.18 ± 0.02	$\mathbf{0.03 \pm 0.002}$	$\mathbf{0.03 \pm 0.002}$

Quantitative results on the synthetic dataset are presented in Table 1. The target registration error (TRE) is defined as the average difference between the ground truth deformation field and the deformation field estimated by the algorithms. DireP was compared to a recent non-rigid registration and PK estimation approach using continuous optimization (Method1) [6]. DireP $3D$ is the algorithm version without temporal regularization and DireP $4D$ is the variant including temporal regularization. The results show that while both DireP $3D$ and DireP $4D$ outperform Method1 and recover a good part of the synthetic motion, DireP $3D$ has a slightly better performance in terms of TRE. This is due to trading a solution that is globally optimal for each volume (DireP $3D$) for a sub-optimal solution on the entire $4D$ graph (DireP $4D$). After applying each method, the PK parameters on the corrected datasets were generated by least-squares fitting. The errors in k_{ep} and v_e are defined as the average absolute difference between the ground truth PK parameters and the parameters obtained on the corrected datasets. These values were calculated on a circular region delineated around the tumour in the pre-contrast image volume (which

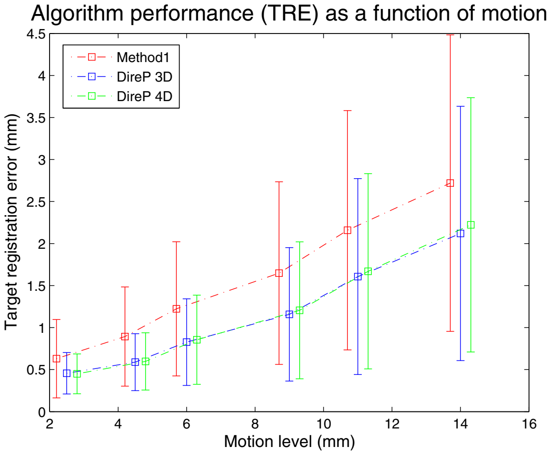


Fig. 2. The TREs (mm) for the three different algorithms are shown in function of the motion level. The performance of Method1 is represented with red error bars, DireP $3D$ is represented in blue, and DireP $3D$ is represented in green. The coloured error bars are slightly displaced for better visualization, but they correspond to the same level of motion.

was taken from a real dataset as explained above). The PK parameter errors decrease considerably after DireP $3D$ and DireP $4D$ registration.

The relationship between the algorithms' performance and the level of motion in the synthetic data was also investigated, and the results are shown in Fig. 2. It can be seen that DireP $3D$ and DireP $4D$ outperform Method1 regardless of the motion level, and the difference is more pronounced for bigger motion.

3.2 Pharmacokinetic Modelling and Motion Correction on DCE-MRI Images of Rectal Cancer

The presented algorithm was also applied to a dataset of DCE-MRI images from 8 patients with advanced rectal tumours. T1-weighted dynamic images of the pelvis were acquired using the spoiled gradient echo LAVA protocol. Contrast agent (ProHance) was injected at a rate of 3ml/sec, 0.1 mmol/kg body weight. An image volume was acquired every 9.5 seconds for approximately 5 minutes, yielding a sequence of $512 \times 512 \times 52 \times 29$ with a resolution of $0.7813 \times 0.7813 \times 2$mm. The algorithms were applied to a $120 \times 120 \times 52 \times 29$ ROI which contained the tumour volume. The tumours were manually delineated by our clinical collaborators. The experiments were run using $n_{iterPK} = 3$, $n_{iterMRF} = 5$, $\gamma_{sp} = 0.01$ and $\gamma_{temp} = 0.1$, which were empirically chosen. For the registration, 4 resolution levels were used. Depending on the resolution level, u was $8, 6, 4, 4$, with a quantization q of $2, 1, 0.5, 0.5$mm. We used groups of voxels of sizes 8^3, 6^3, 4^3 and 2^3, with corresponding label spaces of size 9^3, 7^3, 5^3 and 5^3. The results for the entire clinical dataset are shown in Fig. 3. The

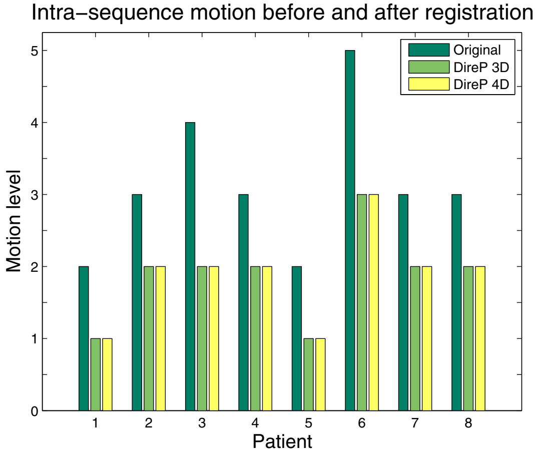


Fig. 3. The DCE-MRI images were blindly graded before and after registration by a clinical expert. A score of '1' represents no motion, '2' is mild or minimal motion, '3' is moderate motion, '4' is significant motion and '5' is severe motion. Both DireP $3D$ and DireP $4D$ reduce the amount of motion in all the images of the rectal cancer dataset.

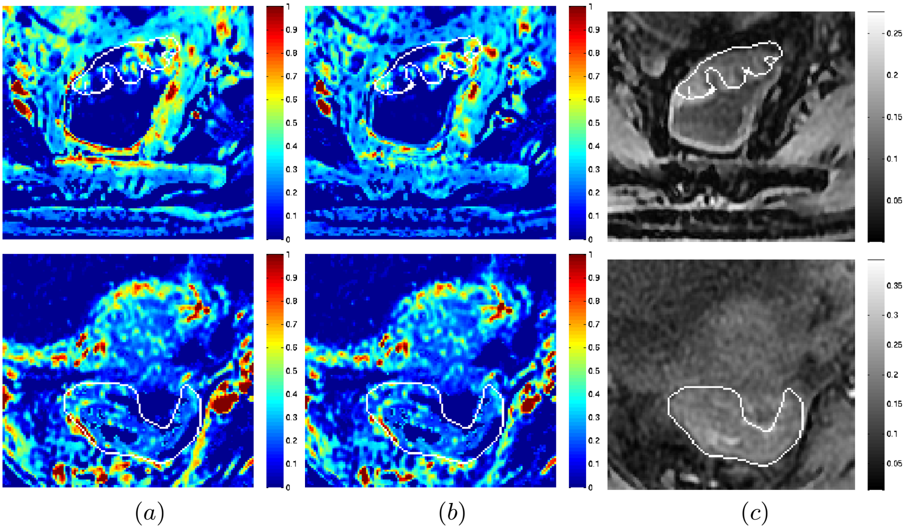


Fig. 4. K_{trans} maps for one slice of the image volume. Images before (a) and after DireP $3D$ (b) motion correction are shown, together with the corresponding anatomical image (c). The top row corresponds to a patient exhibiting severe motion, and the bottom row corresponds to a patient exhibiting moderate motion.

motion level in each time series was blindly graded before and after registration by a clinical expert. It can be seen that the observed motion level decreases in all the patients after applying our discrete registration framework in $3D$ and $4D$, respectively. We also note that the differences between DireP $3D$ and DireP $4D$ were not detectable on a visual evaluation. Figure 4 presents the effect of motion correction on K_{trans}. Images before and after DireP $3D$ are shown, for a patient exhibiting severe motion, and for a patient exhibiting moderate motion. It can be seen that in both cases the K_{trans} maps become sharper after registration, with a clearer separation between individual voxels, and motion artefacts at the tumour boundary are reduced. This effect is particularly visible for the severe motion case, where the non-registered sequence yields a heavily blurred K_{trans} map.

On a 2.93GHz CPU, using C++ code, runtimes are: Method1 14.29min, DireP $3D$ 3.15min, DireP $4D$ 7.02min. For both the $3D$ and the $4D$ algorithm, the Jacobian determinant of the deformation field is positive. The Jacobian determinants were calculated for the real dataset, as well as for the synthetic images.

4 Discussion and Conclusion

We have proposed a new algorithm for DCE-MRI time series motion correction and PK estimation (DireP), which is formulated on an MRF and uses discrete optimization. The PK estimation is performed iteratively with the deformable registration. Two variants of the algorithm, one without temporal regularization (DireP $3D$) and one with temporal regularization (DireP $4D$) were tested on a synthetic dataset and a dataset comprising 8 DCE-MRI sequences of rectal cancer patients. Both DireP $3D$ and DireP $4D$ reduce the amount of motion in all the images of the rectal cancer dataset, especially in challenging sequences exhibiting severe motion. When tested on synthetic data, both the variants show an improvement over a state-of-the art algorithm using continuous optimization [6]. Although DireP $3D$ slightly outperforms DireP $4D$ in terms of TRE ($0.59mm$ vs. $0.62mm$), the key advantage of DireP $4D$ lies in its ability to impose a degree of temporal smoothness, which in itself is desirable to avoid unrealistic fitting of the PK model. At the same time, motions such as peristalsis with a high frequency cannot be captured by a transformation that enforces temporal smoothness. We expect the positive effect of the 4D regularization to be much more visible for applications where lower frequency motion, i.e. breathing, is predominant. Examples include liver DCE-MRI, or rectal data where peristalsis is controlled by drug administration.

The pre-contrast image was chosen as a reference for registration as it more closely represents the true appearance of the anatomy, where features are not distorted by contrast arrival. This also avoids registration error propagation. For the PK modelling part, the standard Tofts model is assumed, as it is widely used in clinical practice. Our algorithm is expected to give comparable if not better results using a more complex model.

5 Future Work

In this work, the $4D$ optimization problem was solved by alternating between spatial MST and temporal chain minimization, and averaging the resulting marginals. This procedure was repeated a couple of times akin LBP [10]. Although this approach has the advantage of reduced complexity, as it involves optimizing loop-free graphs, convergence might be slow and is not guaranteed. In future work, algorithms for joint minimization of the spatial and temporal problem will be investigated. A possible approach is employing a $4D$ generalization of the TRW-S algorithm [11].

Acknowledgements. Monica Enescu is funded from an EPSRC Doctoral Training Award, an Oxford Department of Engineering Science Scholarship and the Alberto Del Vicario Scholarship. Julia Schnabel and Esme Hill gratefully acknowledge funding from CRUK/EPSRC Centre for Cancer Imaging at Oxford. Esme Hill was also funded by the Oxfordshire Health Services (OHSRC). Ricky Sharma acknowledges funding from NIHR BRC Oxford.

References

1. Zoellner, F.G., Sancee, R., Rogelj, P., Ledesma-Carbayo, M.J., Rorvik, J., Santos, A., Lundervold, A.: Assessment of 3D DCE-MRI of the kidneys using non-rigid image registration and segmentation of voxel time courses. Comput. Med. Imag. Graphics 33(3) (2009)
2. Tanner, C., Schnabel, J.A., Chung, D., Clarkson, M.J., Rueckert, D., Hill, D.L.G., Hawkes, D.J.: Volume and shape preservation of enhancing lesions when applying non-rigid registration to a time series of contrast enhancing MR breast images. In: Delp, S.L., DiGoia, A.M., Jaramaz, B. (eds.) MICCAI 2000. LNCS, vol. 1935, pp. 327–337. Springer, Heidelberg (2000)
3. Melbourne, A., Atkinson, D., White, M.J., Collins, D., Leach, M., Hawkes, D.: Registration of dynamic contrast-enhanced MRI using a progressive principal component registration (PPCR). Phys. Med. Biology 52(17) (2007)
4. Hayton, P., Brady, M., Tarassenko, L., Moore, N.: Analysis of dynamic MR breast images using a model of contrast enhancement. Medical image analysis 1(3) (1997)
5. Buonaccorsi, G.A., O'Connor, J.P.B., Caunce, A., Roberts, C., Cheung, S., Watson, Y., Davies, K., Hope, L., Jackson, A., Jayson, G.C., Parker, G.J.M.: Tracer kinetic model-driven registration for dynamic contrast-enhanced MRI time-series data. Magnetic Resonance in Medicine 58(5) (2007)
6. Bhushan, M., Schnabel, J.A., Risser, L., Heinrich, M.P., Brady, J.M., Jenkinson, M.: Motion correction and parameter estimation in dceMRI sequences: Application to colorectal cancer. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part I. LNCS, vol. 6891, pp. 476–483. Springer, Heidelberg (2011)
7. Prim, R.: Shortest connection networks and some generalizations. Bell System Technical Journal (36) (1957)
8. Glocker, B., Komodakis, N., Tziritas, G., Navab, N., Paragios, N.: Dense image registration through MRFs and efficient linear programming. Medical Image Analysis 12(6) (2008)
9. Boykov, V., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Tran. PAMI 23, 1222–1239 (2001)
10. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. Intl. J. Computer Vision 70(1) (2006)

11. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. IEEE Tran. PAMI 28(10) (2006)
12. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: A survey. IEEE Tran. Medical Imaging 32(7) (2013)
13. Tofts, P.S.: Modeling tracer kinetics in dynamic Gd-DTPA MR imaging. J. Magnetic Resonance Imaging 7(1) (1997)
14. Orton, M.R., d'Arcy, J.A., Walker-Samuel, S., Hawkes, D.J., Atkinson, D., Collins, D.J., Leach, M.O.: Computationally efficient vascular input function models for quantitative kinetic modelling using DCE-MRI. Phys. Med. Biology 53(5) (2008)
15. Heinrich, M.P., Jenkinson, M., Brady, M., Schnabel, J.A.: MRF-based deformable registration and ventilation estimation of lung CT. IEEE Tran. Medical Imaging 32(7) (2013)
16. Heinrich, M.P., Simpson, I., Jenkinson, M., Brady, M., Schnabel, J.A.: Uncertainty estimates for improved accuracy of registration-based segmentation propagation using discrete optimization. In: MICCAI SATA Workshop (2013)
17. Felzenszwalb, P.F., Zabih, R.: Dynamic programming and graph algorithms in computer vision. IEEE Tran. PAMI 33(4), 721–740 (2011)
18. Hirschmueller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Tran. PAMI 30(2), 328–341 (2008)
19. Arsigny, V., Commowick, O., Pennec, X., Ayache, N.: A log-euclidean framework for statistics on diffeomorphisms. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 924–931. Springer, Heidelberg (2006)

Learning Imaging Biomarker Trajectories from Noisy Alzheimer's Disease Data Using a Bayesian Multilevel Model

Neil P. Oxtoby[1], Alexandra L. Young[1], Nick C. Fox[2], The Alzheimer's Disease Neuroimaging Initiative*, Pankaj Daga[1], David M. Cash[2,1], Sebastien Ourselin[1], Jonathan M. Schott[2], and Daniel C. Alexander[1]

[1] Progression Of Neurodegenerative Disease Initiative, Centre for Medical Image Computing, Department of Computer Science, University College London, Malet Place, London, WC1E 6BT, UK

[2] Dementia Research Centre, Institute of Neurology, University College London, 8-11 Queen Square, London, WC1N 3AR, UK

Abstract. Characterising the time course of a disease with a protracted incubation period ultimately requires dense longitudinal studies, which can be prohibitively long and expensive. Considering what can be learned in the absence of such data, we estimate cohort-level biomarker trajectories by fitting cross-sectional data to a differential equation model, then integrating the fit. These fits inform our new stochastic differential equation model for synthesising individual-level biomarker trajectories for prognosis support. Our Bayesian multilevel regression model explicitly includes measurement noise estimation to avoid regression dilution bias. Applicable to any disease, here we perform experiments on Alzheimer's disease imaging biomarker data — volumes of regions of interest within the brain. We find that Alzheimer's disease imaging biomarkers are dynamic over timescales from a few years to a few decades.

1 Introduction

Dementia presents a significant societal and economic burden to an ageing population. Late-onset dementia is generally attributed to degenerative neurological diseases such as Alzheimer's disease (AD). Biomarkers (biological markers) are indicators of disease-specific changes which can be used to inform the diagnosis of AD [1]. While no single biomarker is dynamic over the entire disease progression, AD biomarker abnormality is hypothesised to occur in a disease-specific sequence determined by the maximum gradient [2]. Most investigations of this hypothesis have sought to correlate biomarker gradient/change with clinically-determined subject cognition: cognitively normal (CN), mild cognitive impairment (MCI), or diagnosed AD [3]. This approach provides a coarsely-graded ordering of biomarker abnormality. More finely-graded sequencing of biomarker abnormality events has been achieved using data-driven models of disease progression [4–6]. Such models can be useful for diagnosis/staging of a patient, but accurate prognosis requires more complete knowledge of biomarker trajectories.

* ADNI See p. 93.

M.J. Cardoso et al. (Eds.): BAMBI 2014, LNCS 8677, pp. 85–94, 2014.

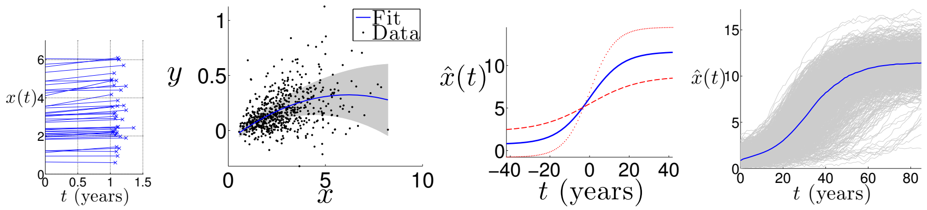
Characterising biomarker dynamics ultimately requires long-term, dense, longitudinal studies. Such data is expensive and difficult to obtain, whereas cross-sectional (or short-term longitudinal) data is relatively inexpensive, easy to obtain, and already available. For example, the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. In this study we present a principled approach to quantitative biomarker dynamics. We start by estimating cohort-level (average) biomarker trajectories by integrating a parametric ordinary differential equation model which is fit to single-followup cross-sectional data, such as done in similar previous work [7–10, 19]. We innovate on previous work in two ways: 1) modelling (and estimating from the data) biomarker measurement noise using a Bayesian multilevel model (BMM); and 2) introducing a stochastic differential equation model for synthesising future biomarker trajectories of individuals, thus providing predictive/prognostic information. We describe the data and methods in Section 2, present results in Section 3, and discuss in Section 4.

2 Data and Methods

From ADNI-1* we consider a cross section of differential data $(x, y \equiv dx/dt)$ for each of five imaging biomarkers (see table 1). Here x is the baseline biomarker value (volume of a region of interest) and y is the forward finite-difference approximation of the derivative from baseline to 12-months. The regional brain volumes are normalised by intracranial volume [11] and presented as percentages. We focus our experimental results on only one region of interest, choosing the ventricles. Results for the other brain volumes are summarised. To maintain specificity to disease progression we included the entire cognitive spectrum except for non-stable or non-progressing individuals (mixed or regressing diagnoses). Excluding individuals with missing data left $N = 651$ individuals.

Illustrative Example of our Approach. Figure 1 illustrates the pipeline of our approach using ventricles data. The single-followup data in figure 1a produces a differential cross section, which is fit to a polynomial differential equation in figure 1b. Integrating the differential equation produces the cohort-level trajectory in figure 1c. The solid blue line shows the average, with dotted and dashed red lines respectively showing short and long transitions from the ±1 standard error bounds on the model parameter estimates. Figure 1d shows individual-level trajectories synthesised by a stochastic differential equation. We proceed now to present details of our methodology.

Regression Model. For each biomarker $x(t)$ we performed model selection using the sample-size-corrected Akaike information criterion. For this purpose we used ordinary least squares (OLS) differential equation models $y = f(x)$, with polynomials $f(x)$ of up to second-order, as well as linear dependence on mean-centred covariates — age and education. We considered group differences by sex, and performed a separate regression for the whole cohort and for the apoE4+ subcohort of genetic risk factor carriers (apoE4 = apolipoprotein-E4). Of the $N = 651$ stable or progressing individuals with suitable brain volumetry data in ADNI-1, 321 were apoE4+ (had one or more apoE4 alleles). Distinct



(a) Data **(b)** Differential fit **(c)** Est. trajectory **(d)** Stochastic model

Fig. 1. Pipeline illustrated on ventricles: (a) single-followup cross-section; (b) differential equation fit; (c) cohort-level trajectory; (d) individual synthetic trajectories (see also figure 3).

from previous work, we use a multilevel differential equation model that incorporates additive Gaussian noise on the biomarker observations $\tilde{x}(t) = x(t) + \eta(t)$. In general, the Gaussian random variable $\eta \sim \mathcal{N}(0, \sigma_{\mathrm{z}}^2)$ may exhibit longitudinal correlation, but since cross-sectional data cannot support estimation of such intra-subject variance, we assume the measurement noise autocorrelation coefficient $\rho(t-s) \equiv \mathrm{E}\left[\eta(t)\eta(s)\right]/\sigma_{\mathrm{z}}^2$ to be $\rho = 0$. (We retain ρ in the covariance matrix below for completeness.)

Our multilevel model has three levels: dynamics (one data point per $i$ndividual), one group level to capture sex differences ($s[i]$), and additive Gaussian measurement noise:

$$y_i \sim \mathcal{N}\left(f(x_i, \boldsymbol{\mu}_{s[i]}), \sigma_{\mathrm{y}}^2\right) \\ \begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} x_i \\ y_i \end{pmatrix}, \Sigma\right) \tag{1}$$

where polynomial $f(x, \boldsymbol{\mu})$ is the dynamical model (see below) parametrised by the vector $\boldsymbol{\mu}$ of sex-specific fixed effects and σ_{y}^2 is residual model error (unexplained variance). The finite-difference derivative $y_i \approx (x_i(T_i) - x_i(0))/T_i$ is correlated with $x_i = x_i(0)$ giving the measurement covariance structure

$$\Sigma_i = \sigma^2 \begin{bmatrix} 1, & -(1-\rho)/T_i \\ -(1-\rho)/T_i, & 2(1-\rho)/T_i^2 \end{bmatrix} \tag{2}$$

where we assume zero intra-subject autocorrelation as discussed above, so $\rho = 0$. The precise value of T_i is used (nominally $T = 1$ year).

Our Bayesian multilevel models are fit numerically using Markov Chain Monte Carlo (MCMC) techniques. For this purpose we used the Stan [12] software package. Full validation is a topic for future work, but it is reassuring to note that we found similar results using the JAGS [13,14] software package. We contrast our results with those obtained using OLS. To ensure that the estimation was driven by the data, our Bayesian models used weakly-informative priors: broad Gaussian priors ($\sigma \geq 100$) for regression parameters and broad positive uniform priors (upper bound $10\tilde{x}_{\max}$) for the variance parameters. We tried different weakly-informative priors (e.g., inverse gamma priors for variances) and the results were unchanged, giving us confidence that the data was driving our inference.

Cohort-level Biomarker Trajectories. Each cohort-level biomarker trajectory is ultimately determined by the (unknown) disease stage for each subject, and represented within our model by the regression parameter estimates. Linear fits produce exponential trajectories, where $f(x) = \mu_{0,s[i]} + \mu_{1,s[i]}x + \mu_{3,s[i]} \cdot \text{age} + \mu_{4,s[i]} \cdot \text{edu}$. This corresponds to acceleration or deceleration/saturation of atrophy in the brain. Quadratic fits, where $f(x) = \mu_{0,s[i]} + \mu_{1,s[i]}x + \mu_{2,s[i]}x^2 + \mu_{3,s[i]} \cdot \text{age} + \mu_{4,s[i]} \cdot \text{edu}$, are a parsimonious representation of sigmoidal trajectories (acceleration followed by deceleration). We note the following convenient analytical form for a sigmoidal trajectory

$$x(t) = x_- + \frac{\varDelta}{1 + e^{-rt}} \tag{3}$$

where r is a biomarker progression rate, and $\varDelta \equiv x_+ - x_-$ separates the asymptotes $x_\pm \equiv x(t = \pm\infty)$ which bound the sigmoid. Time t is symmetric about the primary inflection point $x(t = 0) = x_- + \varDelta/2$, and is unrelated to study/visit time (since baseline). The sigmoid parameters are straightforward functions of the regression parameters $\boldsymbol{\mu}$. We will utilise equation (3) to define a timescale of interest for the cohort (see equation (4) in section 3).

Individual-level Biomarker Trajectories. The ordinary differential equation reflects biomarker dynamics at the cohort level. We model individual-level biomarker dynamics as deviations about this average using a corresponding stochastic differential equation driven by a zero-mean Gaussian process $d\kappa \sim GP(0, \sigma_\kappa^2)$. We propose a prognostic utility for this below.

Biomarker Abnormality Timescales. Model fitting is followed by estimation of a biomarker abnormality timescale for the cohort, and one for individuals. The first is a cohort-level estimate of the duration over which the biomarker is dynamic: between two extremal thresholds $x_\mathrm{s}(t_\mathrm{s})$ (effective *s*aturation) and $x_\mathrm{a}(t_\mathrm{a})$ (initial signs of *a*bnormality). Choosing these thresholds is an open problem. For sigmoidal trajectories we choose analytical thresholds: the points of maximum biomarker acceleration and deceleration. For exponential trajectories (biomarker timescales not presented here) we propose using thresholds of clinical relevance. The second timescale uses our stochastic model to estimate an analogous result for an individual j. Starting at the individual's initial measurements $(\tilde{x}_j, \tilde{y}_j)$, many stochastic trajectories are synthesised using the deviation from the cohort fit as the Gaussian process scale $\sigma_{\kappa,j} = |\tilde{y}_j(\tilde{x}_j) - \hat{y}_\mathrm{fit}(\tilde{x}_j)|$, and sampling model parameters $\boldsymbol{\mu}_j$ from the posterior distributions of the cohort-level multilevel regression parameters. The average of these synthetic trajectories for an individual gives a density of first-passage times (see [15]) taken to reach some maximal threshold, e.g., the effective saturation threshold in the case of biomarker saturation. This is an interval estimate of time remaining until an individual's biomarker becomes fully abnormal, which can inform prognosis either on it's own, or as part of a panel of such times for multiple biomarkers.

3 Results

MCMC. The Bayesian Multilevel Model (BMM) fitting converged using 2 chains, 2000 burn-in samples, and 8000 MCMC samples, thinned by 2. That is, we observed Gelman's potential scale reduction factor [16] to be $PSRF < 1.1$ for all parameters and hyperparameters, as well as observing all Monte Carlo standard errors to be lower than the posterior standard deviations.

Regression. As expected, our BMM produces different results to OLS — for example, the different quadratic regression fits shown for ventricles in figure 2a (males and females were pooled together in this figure). Multilevel regression parameter estimates for $\boldsymbol{\mu}$ are in table 1. The only data supporting a sigmoidal

Table 1. Multilevel regression fit results: mean (±std) $\times 10^{-3}$. ADNI-1 data at baseline and 12 months were available for $N = 651$ (370 male) stable or progressing subjects – of these, 321 (185 male) were apoe4+ subjects. Sex-specific regression parameters are $\mu_{k,s}$ with $k = 0, 1, 2$ (polynomial coefficients) and $s = m, f$ (male,female).

Biomarker, x	$\mu_{0,m}$	$\mu_{0,f}$	$\mu_{1,m}$	$\mu_{1,f}$	$\mu_{2,m}$	$\mu_{2,f}$	σ
Ventricles – all	−126 (52)	−81 (58)	166 (31)	122 (44)	−17 (4)	−6 (8)	9 (4)
– apoE4+	−174 (75)	−76 (94)	214 (42)	123 (73)	−23 (5)	−3 (13)	
Hippocampus – all	−19 (5)	−20 (6)	29 (14)	21 (15)	n/a	n/a	6 (2)
– apoE4+	−17 (8)	−17 (9)	−19 (23)	11 (24)	n/a	n/a	
Entorhinal cortex – all	−0.4 (4)	−10 (4)	−26 (16)	8 (20)	n/a	n/a	2 (1)
– apoE4+	4.9 (5.1)	−13 (7)	−59 (24)	25 (32)	n/a	n/a	
Fusiform – all	−28 (21)	−47 (23)	2 (20)	21 (21)	n/a	n/a	8 (3)
– apoE4+	−13 (30)	−55 (31)	−21 (29)	24 (29)	n/a	n/a	
Mid. temp. gyrus – all	−39 (24)	−56 (28)	7 (20)	22 (23)	n/a	n/a	10 (3)
– apoE4+	−48 (34)	−39 (40)	5 (29)	4 (34)	n/a	n/a	

Table 2. Ordinary least squares fit results: mean (±std) $\times 10^{-3}$. Compare with multilevel regression results in Table 1.

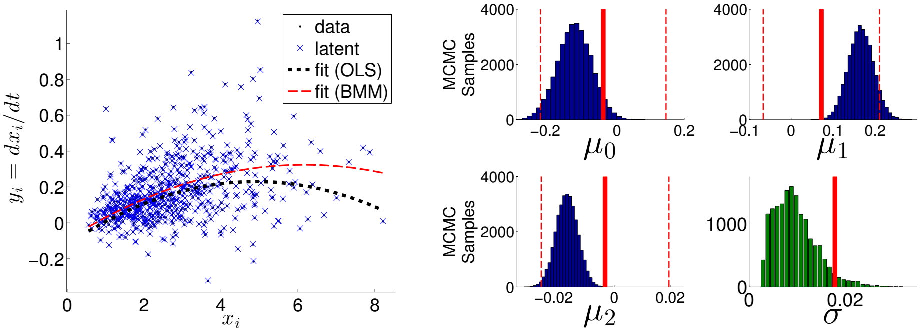
Biomarker, x	μ_0	μ_1	μ_2
Ventricles – all	−41 (62)	72 (46)	−3 (8)
– apoE4+	−92 (96)	130 (72)	−7 (12)
Hippocampus – all	−16 (7)	21 (17)	n/a
– apoE4+	−19 (10)	20 (26)	n/a
Entorhinal cortex – all	−10 (5)	16 (22)	n/a
– apoE4+	−19 (7)	52 (33)	n/a
Fusiform – all	−50 (29)	33 (26)	n/a
– apoE4+	−64 (37)	38 (35)	n/a
Mid. temp. gyrus – all	−16 (33)	0 (27)	n/a
– apoE4+	−22 (45)	−5 (37)	n/a

trajectory was the ventricles of males. The corresponding parameters for equation (3) are shown in Table 3. In both tables, estimates (posterior means) exceeded in magnitude by their standard errors (posterior standard deviations) are effectively zero. From this we can infer the biomarkers for which this combination of data and model implies undetectable change. Acceleration in hippocampal atrophy was detected for the stables/progressors, but not for the apoE4+ subset. Deceleration of atrophy was detectable in the entorhinal cortex of males, but not in females. And for the other regions of interest (fusiform and middle temporal gyrus), this combination of data and model implied undetectable change.

Focussing on ventricles, figure 2b overlays the OLS parameter estimates (vertical lines) upon histograms of the MCMC samples from the BMM. The OLS results differ considerably from the BMM results in value and confidence (spread), resulting in considerably different estimates of the dynamic duration for the biomarker: $\tau = 19 \pm 6$ years (BMM) versus $\tau = 33 \pm 26$ years (OLS). We hypothesize that the BMM has removed a bias present in the OLS regression estimate due to ignoring the measurement noise present in x. The lower right of figure 2b shows a histogram for the measurement noise scale σ from equation (2), compared to the offline estimate from stable controls (red line), which appears to be an overestimate.

We note that there were differences between males and females. The most impressive were for the entorhinal cortex, where the linear differential equation gradients had different signs. Further investigation would require more data and perhaps modelling, so we relegate it to future work.

Our estimates of the Gaussian measurement noise "size" (standard deviation) were all of the order of $\sigma \sim 10^{-3} \approx 0.1\%$ of intracranial volume. This represented between one-third and one-half of the model residual size σ_{y}.



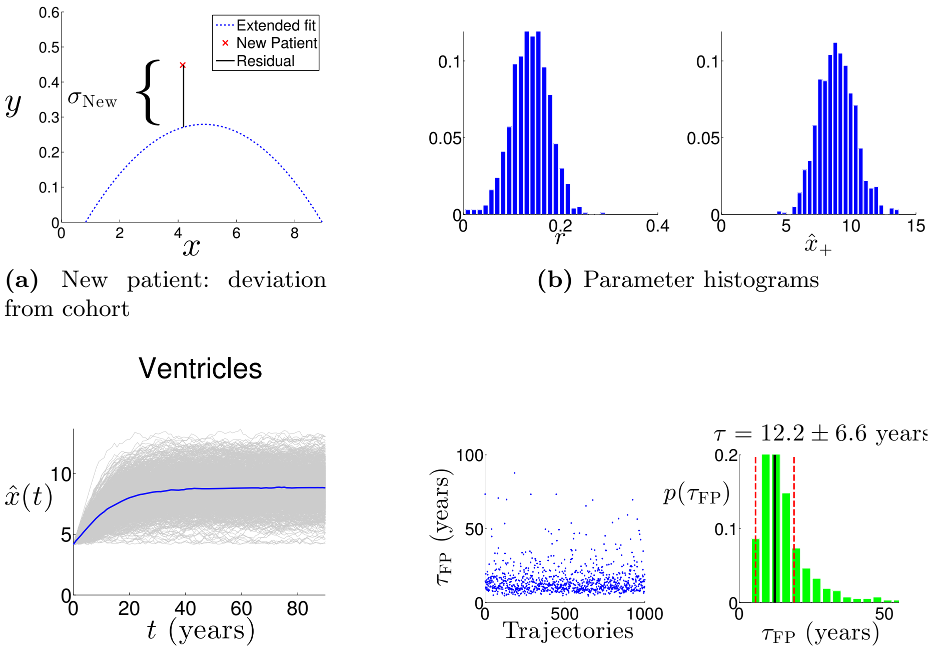
(a) Multilevel model fits

(b) Model parameter estimates

Fig. 2. Regression results for ventricles. Histograms for the multilevel fit parameter MCMC samples are shown with overlays (red lines) of the complete pooling regression results for $\mu_k \pm 3$ standard error. The measurement noise histogram (lower right; green) is compared with the variance in ADNI-1 stable control ventricles measurements, averaged across individuals.

Table 3. Sigmoid parameters and biomarker dynamic duration results for the ventricles of males

Biomarker, x	r, per year	x_-	x_+	x_a	x_s	τ, years
Ventricles (males) – BMM	0.14 (0.04)	0.8 (0.9)	8.9 (1.4)	2.5	7.2	19 (6)
Ventricles (males) – OLS	0.081 (0.065)	0.7 (2.6)	8.8 (4.0)	2.4	7.1	33 (26)



(a) New patient: deviation from cohort

(b) Parameter histograms

(c) Sample paths

(d) First-passage times

Fig. 3. Prognostic utility of our approach: stochastic model. See text for details.

Cohort-level Biomarker Abnormality Timescale. For sigmoidal dynamics, our analytical thresholds for initial abnormality and effective saturation (points of maximal acceleration and deceleration) are found by using equation (3) and solving $dx^3/dt^3 = 0$. The time interval between these thresholds is

$$\tau = \frac{1}{r} \ln \left(\frac{2+\sqrt{3}}{2-\sqrt{3}} \right) . \tag{4}$$

We found $\tau = 19 \pm 6$ years for ventricles in males (the only data to support a sigmoidal trajectory). For exponential biomarker trajectories (not presented here), clinically-relevant thresholds would be appropriate for estimating τ.

Individual-level Biomarker Abnormality Timescale. We calculated the biomarker effective saturation time for the ventricle volume of a randomly-selected individual (RID=1384; diagnosed MCI) at visits not included in the original fit (to avoid circularity): 24 and 36 months. This data point and the resulting residual are shown with the pre-existing cohort fit in Figure 3a. The model parameters sampled from the BMM posterior distributions shown in Figure 3b were used to synthesise the 1000 trajectories in Figure 3c. The corresponding first-passage times are shown in Figure 3d. Due to the long tail of the distribution, we used robust statistics (median ± median absolute deviation) to calculate the biomarker effective saturation time as $\tau_{\mathrm{FP}} = 12.2 \pm 6.6$ years from a ventricular volume of $x_j = 4.1\%$ to the saturation threshold of $x_{\mathrm{s}} = 8.9\%$ (percentage of intracranial volume).

4 Discussion

Neurodegeneration causes the ventricles to expand and all other brain volumes to decline. Measurement noise and intra-subject variability confound this, e.g., some progressing individuals display $y > 0$ even for brain volumes which should be in decline. Indeed, the apparent bias in OLS results suggests that measurement noise should be modelled in a differential equation approach. Quantitatively we found that ventricles saturated after an expansion lasting approximately two decades. This timescale is consistent with current knowledge of Alzheimer's disease, and related work on biomarker trajectories [10, 17, 19].

We found low coefficients of determination $R^2 \leq 0.33$, as in related work [10], implying that a small proportion of the variance in the data was explained by the model. This is not particularly surprising for two reasons: 1) cross-sectional data cannot be used to distinguish between inter-subject and intra-subject variance; and 2) the simplicity of the model compared with the unknown complexity of Alzheimer's disease. For example, the observations in ADNI-1 of hippocampal growth (or ventricular contraction) in diseased subjects could be a consequence of intra-subject variation on the relatively short timescale used to calculate biomarker change (~ 1 year compared to the decades-long incubation period). A first step to reduce the influence of such intra-subject variance (not considered here) would be to use the entire set of followup data from ADNI. Given enough data points per individual, inter-subject variance and heteroscedasticity could be explicitly modelled and estimated. There is hope that dense longitudinal data, as it becomes available, will allow fitting of more complex models that explain the data better.

This study addressed an important problem: how to infer information about disease biomarker trajectories from noisy cross-sectional data, which is readily-available and relatively inexpensive. Cohort-level trajectories were estimated by fitting an ordinary differential equation model, and integrating the fit. Individual-level trajectories were modelled as Gaussian deviations from the cohort using a stochastic differential equation model, allowing trajectory synthesis to inform prognosis. We innovated over previous differential equation models in two ways. First by using a Bayesian multilevel regression model to separately identify measurement noise and population variance. Our second innovation was the stochastic model. The Bayesian multilevel model avoids biassed parameter estimates, which can arise due to regression dilution. Experiments were performed on Alzheimer's disease imaging data from the ADNI. We presented full results only for ventricle volume (a quadratic differential equation with sigmoidal time course), but our framework is not limited to a particular dynamical model.

In conclusion, clinicians focussing on patient outcomes ultimately desire improved diagnosis and prognosis — informed by biomarkers, including those derived from medical image computing. Prognostic uncertainty can be as important to the patient as the prognosis itself [18], so it is crucial to provide interval estimates of relevant time scales where possible. Our stochastic model allows interval estimation of the time remaining until a biomarker approaches maximal abnormality. A panel of such estimates for multiple biomarkers could be used to inform prognosis, e.g., estimation of time until onset of dementia. In the future we envisage developing such a prognostic tool using our approach in concert with disease progression models and/or longitudinal quantitative tools such as recurring-event survival analysis.

*Data used in preparation of this article was obtained from the ADNI database (`adni.loni.usc.edu/`). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at `http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf`. Funding for the ADNI project came from National Institutes of Health Grant U01 AG024904. Further details of funding for the ADNI are available via the ADNI website: `www.adni-info.org`.

References

1. McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack Jr., C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., Mohs, R.C., Morris, J.C., Rossor, M.N., Scheltens, P., Carrillo, M.C., Thies, B., Weintraub, S., Phelps., C.H.: The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia 7(3), 263–269 (2011)
2. Jack, C.R., Knopman, D.S., Jagust, W.J., Shaw, L.M., Aisen, P.S., Weiner, M.W., Petersen, R.C., Trojanowski, J.Q.: Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. The Lancet Neurology 9(1), 119–128 (2010)
3. Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack, C.R., Jagust, W., Liu, E., Morris, J.C., Petersen, R.C., Saykin, A.J., Schmidt, M.E., Shaw, L., Siuciak, J.A., Soares, H., Toga, A.W., Trojanowski, J.Q.: The Alzheimer's Disease Neuroimaging Initiative: A review of papers published since its inception. Alzheimer's & Dementia 8(1), S1–S68 (2012)
4. Fonteijn, H.M., Modat, M., Clarkson, M.J., Barnes, J., Lehmann, M., Hobbs, N.Z., Scahill, R.I., Tabrizi, S.J., Ourselin, S., Fox, N.C., Alexander, D.C.: An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease
5. Huang, J., Alexander, D.: Probabilistic Event Cascades for Alzheimer's disease. In: Advances in Neural Information Processing Systems, pp. 3104–3112 (2012)

6. Young, A.L., Oxtoby, N.P., Daga, P., Cash, D.M., Fox, N.C., Ourselin, S., Schott, J.M., Alexander, D.C.: A data-driven model of biomarker changes in sporadic Alzheimer's disease. Brain 137(9), 2564–2577 (2014), http://brain.oxfordjournals.org/content/137/9/2564
7. Ashford, J.W., Schmitt., F.A.: Modeling the time-course of Alzheimer dementia. Current Psychiatry Reports 3(1), 20–28 (2001)
8. Yang, E., Farnum, M., Lobanov, V., Schultz, T., Raghavan, N., Samtani, M.N., Novak, G., Narayan, V., DiBernardo, A.: Quantifying the Pathophysiological Timeline of Alzheimer's Disease. Journal of Alzheimer's Disease 26(4), 745–753 (2011)
9. Sabuncu, M., Desikan, R., Sepulcre, J., Yeo, B., Liu, H., Schmansky, N., Reuter, M., Weiner, M., Buckner, R., Sperling, R.: The dynamics of cortical and hippocampal atrophy in Alzheimer disease. Archives of Neurology 68(8), 1040 (2011)
10. Villemagne, V.L., Burnham, S., Bourgeat, P., Brown, B., Ellis, K.A., Salvado, O., Szoeke, C., Macaulay, S.L., Martins, R., Maruff, P., Ames, D., Rowe, C.C., Masters, C.L.: Amyloid β deposition and neurodegeneration and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. The Lancet Neurology 12(4), 357–367 (2013)
11. Barnes, J., Ridgway, G.R., Bartlett, J., Henley, S.M.D., Lehmann, M., Hobbs, N., Clarkson, M.J., MacManus, D.G., Ourselin, S., Fox, N.C.: Head size, age and gender adjustment in MRI studies: a necessary nuisance? NeuroImage 53(4), 1244–1255 (2010)
12. Stan Development Team: Technical report, Stan: A C++ Library for Probability and Sampling, Version 2.2 (2014)
13. Plummer., M.: JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In: Proceedings of the 3rd International Workshop on Distributed Statistical Computing (2003)
14. Steyvers, M.: Technical report MATJAGS 1.3, psiexp.ss.uci.edu/research/programs_data/jags
15. Jacobs, K.: Stochastic Processes for Physicists. Cambridge University Press (2010)
16. Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. Statistical Science 7(4), 457–472 (1992)
17. Villain, N., Chételat, G., Grassiot, B., Bourgeat, P., Jones, G., Ellis, K.A., Ames, D., Martins, R.N., Eustache, F., Salvado, O., Masters, C.L., Rowe, C.C., Villemagne, V.L., The AIBL Research Group: Regional dynamics of amyloid-β deposition in healthy elderly, mild cognitive impairment and Alzheimer's disease: a voxelwise PiB–PET longitudinal study. Brain 135(7), 2126–2139 (2013)
18. Smith, A.K., White, D.B., Arnold, R.M.: Uncertainty — the other side of prognosis. New England Journal of Medicine 368(26), 2448–2450 (2013)
19. Jack, C.R., Wiste, H.J., Lesnick, T.G., Weigand, S.D., Knopman, D.S., Vemuri, P., Pankratz, V.S., Senjem, M.L., Gunter, J.L., Mielke, M.M., Lowe, V.J., Boeve, B.F., Petersen, R.C.: Brain β-amyloid load approaches a plateau. Neurology 80(10), 890–896 (2013)

Four Neuroimaging Questions that P-Values Cannot Answer (and Bayesian Analysis Can)

Maxime Taquet, Jurriaan M. Peters, and Simon K. Warfield

Computational Radiology Laboratory, Boston Children's Hospital,
Harvard Medical School, USA

Abstract. Null Hypothesis Significance Testing (NHST) is used pervasively in neuroimaging studies, despite its known limitations. Recent critiques to these tests have mostly focused on technical issues with multiple comparisons and difficulties in interpreting p-values. While these critiques are valuable, we believe that they overlook the fundamental flaws of NHST in answering research questions. In this paper, we review major limitations inherent to NHST that we formulate as four research questions insoluble with p-values. We demonstrate how, in theory, Bayesian approaches can provide answers to such questions. We discuss the implications of these questions as well as the practicalities of such approaches in neuroimaging.

1 Introduction

The finding that statistically significant fMRI signal change can be mistakingly observed in a dead salmon performing a mentalizing task [1] and the account of too-high-to-be-true correlations between self-reported behavioral measures and brain activations [2] sparked an heated debate in the brain imaging community about the statistical practice employed in such studies [3,4]. Up to very few exceptions (e.g., [5]), this debate has focused on publication bias, appropriate corrections for multiple comparisons, and reporting of findings in good faith, thereby joining the broader discussion on flawed scientific standards [6]. While defining and advocating good scientific practice is of paramount importance, we feel that this discussion has often diverted the attention from the inappropriateness of null-hypothesis significance testing (NHST) in answering research questions —no matter how meticulously applied. There are indeed important neuroimaging research questions to which NHST provides misleading, if any, answers. In this paper, we review four such questions and we describe how Bayesian methods alleviate the fallacies of NHST. Section 2 recalls the rationale behind NHST and p-values. Section 3-6 review four questions insoluble with p-values. Section 7 discusses the implications of these questions for the neuroimaging community.

2 Brain Imaging Is Free of Type I and Type II Errors

NHST proceeds as a proof by contradiction. If our data are incompatible with some hypothesis, then the hypothesis must be wrong. In empirical science, such

M.J. Cardoso et al. (Eds.): BAMBI 2014, LNCS 8677, pp. 95–106, 2014.

as neuroimaging, a definite claim of incompatibility cannot be achieved and we therefore compute the probability to observe the data (or something even less compatible with the hypothesis), should the hypothesis be true. This probability is called the *p-value* and the hypothesis is called the *null hypothesis*. If the p-value is small enough, then the null hypothesis is rejected with confidence. If, notwithstanding the small p-value, the null hypothesis was actually true, then one makes a Type I error (rejecting a true null hypothesis).

Imagine that we are interested in the interaction between cinephilia (a passionate interest in cinema) and hippocampal volume. We define a null hypothesis (that cinephiles have on average exactly the same hippocampal volume as control subjects), collect MRI data, compute the tissue volume and attempt to refute the null hypothesis. Now, before we endeavor to do so, we may ponder the odds that the null hypothesis is actually true. This probability most likely equals 0%. Cinephiles tend to enjoy more esoteric movies typically played in smaller theaters for which signs are not displayed in the city. This may require cinephiles to develop better spatial navigation skills, which are associated with larger hippocampi [7]. Whether this line of reasoning prevails in the global association between cinephilia and the volume of hippocampi or other opposite effects play a more important role, the probability that there is absolutely no effect of cinephilia on hippocampal volume is essentially null. The upside of this fact is that most researchers in neuroimaging make no Type I errors nor Type II errors (since null hypotheses are always wrong). The downside of it, however, is that the conclusions of NHST in this context are fairly useless, since the null hypothesis can be rejected prior to acquiring any data. Making no Type I nor Type II errors in brain imaging does not imply that we do not make any error. Our errors pertain to the sign and the magnitude of our conclusions, coined Type S and Type M errors by Gelman *et al.* [8,9].

3 Type S Errors: How Confident Are We That Our Finding Is not Opposite to the Truth?

Let us assume that we want to compare the brain connectivity between patients with autism and controls. After comparing the groups, we find out that patients with autism have, on average, significantly weaker connections in the language system ($p < 0.05$). Given this statistically significant result, what are the chances that patients with autism actually have stronger connections in the language system (*i.e.*, what are the odds that my finding is opposite to the truth)? The answer is "we really don't know". To understand why, let us formalize the problem[1]. Let θ_1 and θ_2 be the true mean connectivity in the language system of patients with autism and controls respectively. We assume that $\theta_1, \theta_2 \in [-\infty, \infty]$. Let y_1 and y_2 be the observed mean connectivity in patients with autism and in controls. Assuming normality and equal variance (σ^2) in both groups, we have:

[1] This formalization is greatly inspired from the formalism in [8] that we adapt to better reflect the situation of image-based population studies of the brain, in which more information is available a priori for control subjects than for patients.

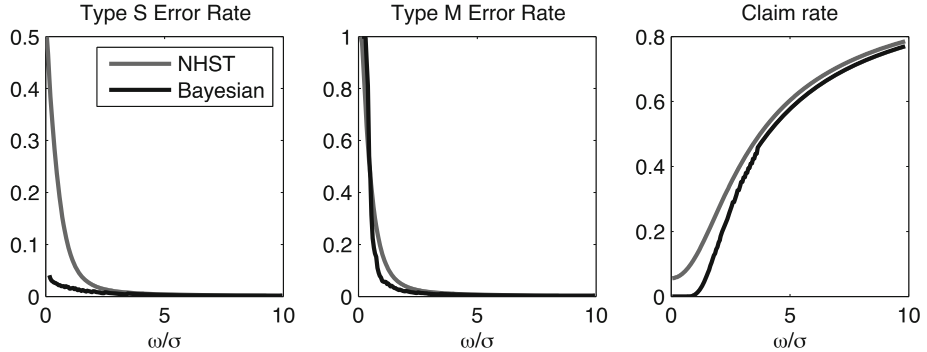


Fig. 1. (Left) NHST makes up to 40% Type S errors for small values of the ratio ω/σ. By contrast, Bayesian analysis controls the Type S error rate to remain below 2.5%. (Middle) Both NHST and Bayesian approaches can make up to 100% Type M errors for small values of ω/σ. (Right) However, for such small values of the ratio, Bayesian approaches are much more prudent than NHST, making very few claims with confidence and showing adaptability to the data as ω/σ increases.

$$y_i|\theta_i, \sigma \sim \mathcal{N}(\theta_i, \sigma^2), \quad i = 1, 2. \tag{1}$$

Type S errors occur whenever $y_1 - y_2$ reaches a specific threshold T to make a claim with confidence (e.g., $|y_1 - y_2| > 1.96\sqrt{2}\sigma$ corresponding to $p < 0.05$) while having a sign that is opposite to the true difference $\theta_1 - \theta_2$. The probability of making a Type S error is therefore given by [8]:

$$P\Big(\text{Type S error}\Big) = P\left(\text{sign}(y_1 - y_2) \neq \text{sign}(\theta_1 - \theta_2)\Big| |y_1 - y_2| > T\right). \tag{2}$$

This probability involves computing the posterior probability over the latent variables θ_i and can therefore only be estimated in a Bayesian approach. Now, although this probability cannot be directly estimated in NHST, is the p-value returned by the test a good enough proxy to the Type S error rate? The answer, as we describe below, is “No”.

To estimate the probability in (2), let us define a simple hierarchical model as in [8]. We assume that σ can reliably be estimated from the data. The prior on θ_1 and θ_2, $p(\theta_1, \theta_2)$, can be expressed conditionally: $p(\theta_1)p(\theta_2|\theta_1)$. The prior $p(\theta_1)$ encodes any prior knowledge that we have about the mean connectivity in controls, as gained, for example, from past experience with such connectivity measures. We may, for instance, assign a normal prior to θ_1 centered at some reasonable μ and some standard deviation τ (our conclusions, as we will see, depend neither on the value μ nor on that of τ):

$$\theta_1|\mu, \tau \sim \mathcal{N}(\mu, \tau^2). \tag{3}$$

If we knew the true value of the mean connectivity in controls, θ_1, and we had no data from patients, our best guess about the value in patients, θ_2, would be

θ_1. We can therefore model the conditional prior on θ_2 as a distribution centered on θ_1, for example a normal with unknown variance ω^2 (we will come back to estimations of the value of ω):

$$\theta_2 | \theta_1, \omega \sim \mathcal{N}(\theta_1, \omega^2). \tag{4}$$

From the hierarchical model described by (1),(3) and (4), we can derive the posterior probability of $\delta \triangleq \theta_1 - \theta_2$ given $d \triangleq y_1 - y_2$ and the joint probability of δ and d:

$$\delta | d, \omega, \sigma \sim \mathcal{N}\left(\frac{d}{1+\frac{2\sigma^2}{\omega^2}}, \frac{1}{\frac{1}{2\sigma^2}+\frac{1}{\omega^2}}\right) \tag{5}$$

$$[d, \delta] | \omega, \sigma \sim \mathcal{N}\left(\mathbf{0}, \omega^2 \begin{pmatrix} 1+\frac{2\sigma^2}{\omega^2} & 1 \\ 1 & 1 \end{pmatrix}\right). \tag{6}$$

Equation (5) allows us to define a 95% posterior interval on δ and therefore define a threshold T_B to make a claim with confidence about the sign of the difference:

$$T_B = 1.96\sqrt{2}\sigma\sqrt{1+\frac{2\sigma^2}{\omega^2}}. \tag{7}$$

Equation (6) then allows us to estimate the Type S error rate for a given threshold T by conditioning on the fact that $|d| > T$ [8]:

$$P\Big(\text{Type S error}\Big) = \frac{\displaystyle\int_{-\infty}^{0}\int_{T}^{\infty} p([d,\delta]|\omega,\sigma)dd\,d\delta + \int_{0}^{\infty}\int_{-\infty}^{-T} p([d,\delta]|\omega,\sigma)dd\,d\delta}{\displaystyle\int_{-\infty}^{+\infty}\int_{T}^{\infty} p([d,\delta]|\omega,\sigma)dd\,d\delta + \int_{-\infty}^{\infty}\int_{-\infty}^{-T} p([d,\delta]|\omega,\sigma)dd\,d\delta}.$$

The conditioning on $|d| > T$ means that we consider a Type S error only if we make an incorrect claim with confidence. This error rate only depends on ω/σ and is depicted in Fig. 1 for $T_F = 1.96\sqrt{2}\sigma$ corresponding to $p < 0.05$ in NHST and for T_B given in Equation (7). For values of $\omega \gg \sigma$, $T_B \approx T_F$ so that both inferences lead to similar Type S errors. However, for $\omega \lesssim \sigma$, the Type S error rate with NHST grows quickly and reaches 40% for $\omega = 0.15\sigma$, demonstrating that we really don't know the odds of making a Type S error given some p-value. By contrast, Type S error rates with the Bayesian approach remain below 5% for all values of ω and is therefore under control.

The ratio ω/σ encodes how far apart we expect, a priori, the variables θ_1 and θ_2 to be with respect to the variance of observations y_1 and y_2. In other words, this ratio encodes our prior on the true underlying effect size and we have:

$$E\left[\left(\frac{\delta}{\sigma}\right)^2\right] = \left(\frac{\omega}{\sigma}\right)^2.$$

NHST can therefore be interpreted as a Bayesian approach which assumes that, a priori, effect sizes are infinite on average. This assumption seems unreasonable and would lead to the observation of extreme group differences in almost all cases. This explains why the actual Bayesian approach performs better for all finite values of ω/σ (and equivalently to NHST for infinite values of ω/σ) as shown on Fig. 1. The actual value of ω could be estimated by pooling all comparisons (all connections) made between the brain of controls and patients. This first example illustrates that NHST cannot resolve important aspects of inference in population studies whereas Bayesian inference enables more adaptive and reliable analyses of the data at hand.

4 Type M Errors: Can the True Effect Be Much Smaller than What We Observed?

Suppose that we are confident (for some ad-hoc reason) that our finding is not a Type S error. Since we never make any Type I and Type II error in brain imaging, what else can invalidate our finding? We may have observed too strong an effect compared to the true effect. This would be a Type M error [9]: the sign of the observed effect may be correct but its magnitude is not.

Again, the question of the prevalence of such errors in practice cannot be answered using NHST alone but can be answered in a Bayesian fashion. Using the same hierarchal model as in the previous section (Equations (1), (3) and (4)), and the resulting posterior and joint distributions (Equations (5) and (6)), we can estimate the Type M error rate. If we define a Type M error as misestimating the effect by a factor 10 ($|d| < |\delta|/10$ or $|d| > 10|\delta|$) while claiming this effect with confidence (*i.e.*, conditionally on $|d| > T$), then the Type M error rate is given by:

$$P\Big(\text{Type M error}\Big) = \frac{\displaystyle\int_T^\infty \int_{\substack{[-\infty,\frac{d}{10}]\\ \cup[10d,\infty]}} p([d,\delta]|\omega,\sigma)d\delta\, dd + \int_{-\infty}^{-T} \int_{\substack{[-\infty,10d]\\ \cup[\frac{d}{10},\infty]}} p([d,\delta]|\omega,\sigma)d\delta\, dd}{\displaystyle\int_{-\infty}^{+\infty} \int_T^\infty p([d,\delta]|\omega,\sigma)dd\, d\delta + \int_{-\infty}^{\infty} \int_{-\infty}^{-T} p([d,\delta]|\omega,\sigma)dd\, d\delta}.$$

The threshold T used in this equation depends on the inference approach being used. For NHST, the conventional 95% confidence interval gives rise to $T_F = 1.96\sqrt{2}\sigma$, whereas the Bayesian approach leads to a 95% posterior interval governed by the threshold T_B of Equation (7). For these thresholds, the Type M error rates are illustrated in Fig. 1. Interestingly, unlike Type S error rates, Type M error rates reach high levels for both NHST and Bayesian approaches for small values of ω/σ. When this ratio falls under 0.45, Type M error rates for both Bayesian and NHST are above 50%, implying that every other finding has an effect that is at least an order of magnitude off compared to the true effect. Not surprisingly, overestimation errors (that is $|d| > 10|\delta|$) are overwhelmingly more present than underestimations in this range (for $\omega/\sigma < 0.5$, approximately all Type M errors consists of overestimations).

Given that Type M errors can reach dramatically high effects with both NHST and Bayesian approaches, one may question what we have gained from the Bayesian approach. There are three reasons for which the Bayesian approach remains beneficial in this case. First, without a Bayesian approach, we would not have been aware of the prevalence of Type M errors, the estimation of which requires the introduction of a prior over latent variables θ_i. Second, for $\omega/\sigma > 0.5$, Type M error rates are consistently smaller with the Bayesian approach than with NHST. Third, and most importantly, we have defined Type M error rates conditionally on making a claim with confidence. Since the threshold for confidence differs between NHST and the Bayesian approach, so will the number of claims being made with confidence. Fig. 1 depicts the rate of claims. Strikingly, the Bayesian approach makes almost no claim with confidence for small values of ω/σ whereas NHST makes at least 5% in all cases (as a consequence of the definition of the p-value). As the ratio ω/σ increases, the number of claims made with the Bayesian approach increases to become closer to the number of claims made with NHST. In other words, the Bayesian approach is always more conservative than NHST (since $T_B > T_F$) and is even more conservative –and rightly so– when the claims will likely lead to a Type M error. This finding shows that, by zealously controlling for hypothetical Type I errors, NHST makes substantially more actual Type M errors. On the other hand, by properly modeling the uncertainty of observations and priors on effect sizes, Bayesian approaches adopt an adaptive behavior in which fewer claims are made with confidence when the data does not justify them.

5 Do Patients and Controls Have Similar Brains?

The probability that two groups of individuals have exactly the same average brain is most often zero. That is because statistics with a continuous domain (for example, the hippocampal volume) often have no mass at zero. In other words, this is because the integral between zero and zero of a finite function is zero. Yet, we do not expect the brains of all patients in all diseases to be affected in all its properties and in all its locations. We expect to observe some *similarities* between brains. But what does *similar* mean in a world where null hypotheses are intrinsically impossible? As we shall see, the answer to this question is not so much statistical as it is biomedical.

First, let us recall why large p-values are no evidence that brains are similar despite its occasional use as such in population studies. Take a somehow well-established neurological finding, for example that patients with agenesis of the corpus callosum (AgCC, the complete or partial absence of a corpus callosum) have disrupted functional inter-hemispheric connections. Now, recruit patients with AgCC and healthy controls, disregard their gender, age and ethnicity, acquire fMRI images, align images to an atlas based on rigid registration only and perform subsequent processing using a version of SPM99 in which a grad student of your lab mistakingly introduced some bugs. In that scenario, the odds of getting a p-value larger than 0.05 are approximately 95% despite the fact that there actually is a true substantial difference between the groups. Furthermore, very large p-values (e.g., $p > 0.9$) occur randomly if the difference between groups

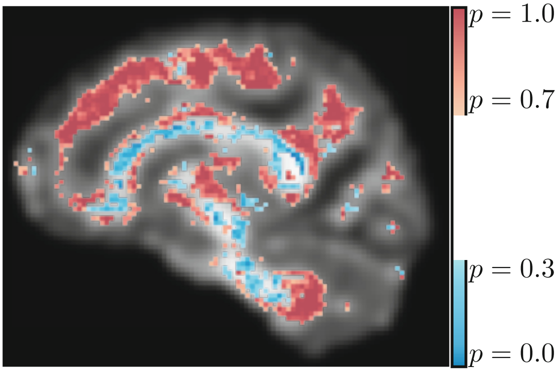


Fig. 2. Example of a map of the posterior probability that the feature of interest (here, the radial diffusivity from [10]) is out of the ROPE: large values are evidence that there is an important difference between the groups (red areas) whereas small values are evidence that the groups are similar (blue areas). The latter cannot be observed in NHST since p-values in those areas are uniformly distributed (and not specially high).

is very little (under the extreme case of zero effect, p-values are uniformly distributed). We therefore cannot increase the threshold on p in a hope to better detect similarities: setting a threshold at $p > 0.95$ would result in at least 95% missed detections of similarities.

If similarities do not imply zero effect, what do they imply? We submit that brains are similar if the difference between them falls within a region of practical equivalence (ROPE) [11] as defined in a Bayesian context in [12]. We therefore want to estimate the probability that the group difference $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$ is within the ROPE:

$$P((\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \in \text{ROPE}|\text{Data})$$

This probability is computed from the posterior distribution over the latent variables and is therefore only computed in a Bayesian framework. Large values of this posterior are evidence that brains are similar between the groups up-to a difference that is within the ROPE (Fig. 2). The definition of the ROPE is not, however, informed by any statistical argument and should rather be answered in a biomedical context. We propose two avenues to define the ROPE: a literature-based based and an introspection-based approach.

Literature-based ROPE. We can establish the ROPE based on other existing studies. For instance, if we want to compare the volume of the hippocampus of patients suffering Alzheimer's disease, we may want to consider that, training for a year as a taxi driver already changes the hippocampal volume by approximately 1% [7]. Since we expect more dramatic changes to occur in the brain of patients with Alzheimer's disease than in the brain of taxi drivers, we may consider brains that differ in their hippocampal volume by up to 1% to be practically equivalent. For studies in pediatrics, one particularly interesting condition to establish relevant orders of magnitude is *age*. If we have curves for the evolution of different brain properties as children develop, we may consider a difference corresponding to one week, one month or one year of maturation as being within the ROPE.

Introspection-based ROPE. In the absence of previous relevant studies, one may ponder the embarrassment involved should the actual magnitude of a statistically significant difference be reported in addition to the p-value. In neuroimaging, p-values are sometimes reported without the actual magnitude of the effect. Imagine that, while measuring the volume of the hippocampus, we observe a difference between controls and patients that is tiny (e.g., 0.06mm^3) yet statistically significant. Would we feel comfortable reporting such a tiny effect? If the answer is "No", then we must probably consider such a difference as belonging to the ROPE. This strategy was used in [10] to set ROPE to microstructural differences in the brain (difference in fascicle directions, diffusivities and volumetric fractions).

6 What Is the Probability That the Patient Has the Disease?

One goal of the definition of biomarkers through population studies is to assist physicians in the decision-making process. In this process, brain images only constitute part of the available information. When making a diagnosis, physicians start off with such information as the patient's age, gender, history and clinical assessment. For instance, the same observed abnormality of a patient's hippocampus is not as strong a sign of Alzheimer's disease in a patient who is 55 years old as it is in a patient who is 68 years old. The odds for the patient to have the disorder also increase if the hippocampal abnormality co-occurs with a clinical presentation of memory impairment or with the presence of a first-degree relative with the disorder. How can all this information be used to estimate the probability that the patient has the disease?

NHST would proceed by eliminating all possible null hypotheses (the null hypothesis that the patient has no disease, then all other null hypotheses that the patient has any other kinds of dementia). After contradicting all null hypotheses, we may believe that Alzheimer's disease is the only possible hypothesis that holds, and yet we would not have any idea of the probability of it being true. This is akin to a diagnosis of exclusion, used in medical practice when no direct conclusive diagnosis can be made.

In the Bayesian formalism, the probability of the disease (D) given all pieces of information can be computed as a posterior probability from the likelihood of the brain imaging data (B) and the prior of the disease given clinical (C) and other individual data (I):

$$\begin{aligned} P(D|B,C,I) &= \frac{P(B|D,C,I)P(D|C,I)}{P(B|C,I)} \\ &= \frac{P(B|D,C,I)P(C|D,I)P(D|I)}{P(B|C,I)P(C|I)} \\ &\propto P(B|D,C,I)P(C|D,I)P(D|I) \end{aligned} \tag{8}$$

Since the denominator does not depend on D and since D is a binary variable, it is sufficient to compute the numerator for both $D = 1$ (has the disease) and $D = 0$ (does not have the disease) and to infer the denominator from the fact that $P(D = 1|B, C, I) + P(D = 0|B, C, I) = 1$. The factor $P(B|D, C, I)$ can be inferred from a model of some kind (for example a generalized linear model), $P(C|D, I)$ can be inferred by calibrating the clinical assessments protocols (for example, those of DSM-V) and $P(D|I)$ is the prevalence of the disease and can usually be obtained from large public health surveys.

Equation (8) stands as a theoretical framework to infer the probability of a patient having the disease given her brain images, her personal and historical information and her clinical assessment. This framework relies on a hierarchical Bayesian model in which prior information can naturally be integrated.

7 Discussion

Throughout the last four sections, we demonstrated that p-values are not appropriate to answer some important brain imaging questions. They fail to predict Type S and Type M error rates (which depend on the prior over effect sizes that is assumed infinite in NHST), they fail to provide evidence for similarities between brains and they cannot be used to estimate the probability that a patient has a disease. We have shown how Bayesian hierarchical models naturally answer those questions. In this section, we discuss the practical implications of these considerations for the neuroimaging research community.

What does Bayesian analysis tell us about the appropriate sample size?. Intuitively, most researchers would probably agree that the more data we have the better the inference. Yet this idea was challenged by Friston in his *Ten Ironic Rules for Non-Statistical Reviewers* [13] on the ground that, for a constant p-value, a significant finding based on fewer samples implies a larger effect size. In t-tests for instance, the p-value is a strictly decreasing function of $d/\sigma = \sqrt{n}d/\sigma'$ where $\sigma' = \sqrt{n}\sigma$ is the standard deviation of the individual samples whereas σ is the standard deviation of their mean in each group (denoted y_1 and y_2 in Section 3 and 4). Increasing n while keeping the p-value constant thus implies decreasing d and therefore decreasing the effect size d/σ'.

However, lower n implies higher σ (since $\sigma = \sqrt{n}\sigma'$ and σ' is determined by the measurement noise and inter-subject variability) which implies lower ω/σ ratios. At lower ω/σ ratios, the Type S and Type M error rates are dramatically higher so that the inference is less reliable as described in Sections 3 and 4. What Friston describes as *larger effect sizes* should be understood as *larger observed effect sizes* which may actually correspond to a *smaller true effect size* likely affected by a Type M error (or even a Type S error), an effect known as p-value filter bias [9] or inflated early-effect sizes [14].

Acknowledging this fallacy of classical inference, Friston further argues that there ought to be a compromise in the choice of a sample size [13]. Too small a sample size would likely lead to inflated early effect-sizes, whereas too high a sample size would result in the detection of trivial effects (*i.e.*, statistically significant effects that are too small to be interesting). Such effects can naturally be accounted for in a Bayesian framework by considering them as practically equivalent to no effect (as described in Section 5). The more data we acquire, the more confident we are that the effect is within or out-of the ROPE and the more accurate our estimate of its sign (fewer Type S errors) and its magnitude (fewer Type M errors). Our Bayesian account of this question therefore indicates that indeed the more data we have, the better.

Bayesian or frequentist inference?. This paper should not be understood as a critique of frequentist inference as a whole. We rather question the appropriateness of NHST in a non-dichotomous context such as brain imaging, much like other researchers have questioned it before in other contexts such as political science [15]. When zero-effects never actually occur, we believe that there is no reason to try hard to control for Type I error rates. Our disagreement with the rationale of NHST can equally be expressed for Bayesian dichotomous analyses such as Bayesian $t-$tests [16] that provide mechanisms to "accept or reject the null hypothesis". In contrast, we believe that there may be some purely frequentist approaches (such as a classifier learned and validated by bootstrapping) that may be appropriate to draw interesting conclusions from brain imaging data, including, for example, to answer the question in Section 6.

Bayesian analysis of neuroimaging data. We believe that the natural answers brought by Bayesian approaches to the four research questions presented above should encourage the neuroimaging community to develop novel Bayesian inference methods. The development of such methods comes with their share of hurdles to overcome. These difficulties pertain to the need for a balance between accurate representation of the data and computational tractability [17]. The challenges in representing brain imaging data as a tractable hierarchical model arises from the *within-voxel* complexity of variables and *between-voxel* dependencies between them.

The increasing complexity of the information contained *within* each voxel leads variables that often belong to non-trivial spaces and with non-trivial dependencies between them. For example, in microstructure imaging, each voxel contains a complete model of the brain microstructure that may present with ten or more variables. These variables belong to the sphere (for directions), the space of strictly positive numbers (for diffusivities and dispersion coefficients) and the simplex (for signal fractions) [10]. A Bayesian hierarchical model should account for these particular spaces. Other examples include the time series contained in fMRI voxels in which contiguous time points are dependent in a non-trivial manner.

Brain imaging variables are also statistically dependent *between* neighboring voxels. In theory, this dependency can be captured by a graphical model, such as a discrete Markov random field [18]. However, graphical models substantially increase the computational complexity when estimating posterior probabilities since the inference needs to be done globally instead of voxel-wise. Typically, in those cases, approximations such as Variational Bayes (VB) methods are employed instead of sampling strategies such as Markov Chain Monte Carlo (MCMC) [18]. These approximations have been shown to outperform non-Bayesian introductions of spatial information in problems such as image registration [19]. However, they are known to introduce biases in the estimations of posterior probabilities [20]. These biases may be of little concern when the inference results of interest are specific values of some variables (e.g., the prior components and the deformation field in [19]) because these values may be exact even when the posterior probability estimate is not. Biases in estimates of the posterior may be more concerning when the goal of inference is to obtain the actual value of the posterior probability, as when assessing the probability that a patient has some disease. In those cases, MCMC sampling and its computational burden may be unavoidable or the bias caused by VB methods should be proved negligible. We believe that these are important avenues for future research.

8 Conclusion

Null hypothesis significance testing (NHST) is a well-defined concept that, if properly conducted, results in a mathematically correct p-value. This p-value, in neuroimaging, is however often useless since all null hypotheses can readily be refuted on the basis that any condition affects our brain in some way. There are therefore many important research questions in neuroimaging that NHST cannot answer. This paper has reviewed these questions and proposed Bayesian alternatives to answer them. Bayesian approaches reduce inference errors (Type S and Type M), enable the building of evidence for the presence of similarities between brains and the incorporation of prior information in the diagnosis, as gleaned from clinical history and examinations. To leverage Bayesian approaches in neuroimaging analyses, technical difficulties related to the complexity of the information within and between voxels must be overcome. Important methodological developments in this area are being made and should be sustained and expanded to move the neuroimaging community away from the inappropriateness of NHST.

Acknowledgments. This work was supported in part by WBI, NIH grants R01 NS079788, R01 LM010033, R01 EB013248, P30 HD018655, and by a research grant from the Boston Children's Hospital Translational Research Program.

References

1. Bennett, C.M., Miller, M., Wolford, G.: Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for multiple comparisons correction. In: Organization for Human Brain Mapping, pp. S39–S41 (2009)
2. Vul, E., Harris, C., Winkielman, P., Pashler, H.: Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. Perspectives on psychological science 4(3), 274–290 (2009)
3. Lieberman, M.D., Berkman, E.T., Wager, T.D.: Correlations in social neuroscience aren't voodoo: commentary on Vul et al. Perspectives on Psychological Science 4(3), 299–307 (2009)
4. Nichols, T.E.: Multiple testing corrections, nonparametric methods, and random field theory. Neuroimage 62(2), 811–815 (2012)
5. Lindquist, M.A., Gelman, A.: Correlations and multiple comparisons in functional imaging: a statistical perspective (commentary on vul et al., 2009). Perspectives on Psychological Science 4(3), 310–313 (2009)
6. Ioannidis, J.P.: Why most published research findings are false. PLoS Medicine 2(8), e124 (2005)
7. Maguire, E.A., et al.: Navigation-related structural change in the hippocampi of taxi drivers. Proc. Nat. Acad. Sci. 97(8), 4398–4403 (2000)
8. Gelman, A., Tuerlinckx, F.: Type S error rates for classical and bayesian single and multiple comparison procedures. Computational Statistics 15(3), 373–390 (2000)
9. Gelman, A., Weakliem, D.: Of beauty, sex and power. American Scientist 97(4), 310–316 (2009)
10. Taquet, M., Scherrer, B., Peters, J.M., Prabhu, S.P., Warfield, S.K.: A fully bayesian inference framework for population studies of the brain microstructure. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part I. LNCS, vol. 8673, pp. 25–32. Springer, Heidelberg (2014)
11. Jones, B., Jarvis, P., Lewis, J., Ebbutt, A.: Trials to assess equivalence: the importance of rigorous methods. BMJ 313(7048), 36–39 (1996)
12. Kruschke, J.K.: Bayesian assessment of null values via parameter estimation and model comparison. Perspectives on Psychological Science 6(3), 299–312 (2011)
13. Friston, K.: Ten ironic rules for non-statistical reviewers. Neuroimage 61(4), 1300–1310 (2012)
14. Ioannidis, J.P.: Why most discovered true associations are inflated. Epidemiology 19(5), 640–648 (2008)
15. Gelman, A.: Commentary: p values and statistical practice. Epidemiology 24(1), 69–72 (2013)
16. Rouder, J.N., Speckman, P.L., Sun, D., Morey, R.D., Iverson, G.: Bayesian t tests for accepting and rejecting the null hypothesis. Psychonomic Bulletin & Review 16(2), 225–237 (2009)
17. Gelman, A., Shalizi, C.R.: Philosophy and the practice of bayesian statistics. British Journal of Mathematical and Statistical Psychology 66(1), 8–38 (2013)
18. Woolrich, M.W., et al.: Bayesian analysis of neuroimaging data in FSL. Neuroimage 45(1), S173–S186 (2009)
19. Simpson, I.J.A., Woolrich, M.W., Cardoso, M.J., Cash, D.M., Modat, M., Schnabel, J.A., Ourselin, S.: A bayesian approach for spatially adaptive regularisation in non-rigid registration. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 10–18. Springer, Heidelberg (2013)
20. Hoffman, M.D., Gelman, A.: The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. arXiv preprint arXiv:1111.4246 (2011)

Spherical Topic Models for Imaging Phenotype Discovery in Genetic Studies

Kayhan N. Batmanghelich[1], Michael Cho[2],
Raul San Jose[2], and Polina Golland[1]

[1] Computer Science and Artificial Intelligence Lab., MIT, USA
[2] Brigham and Women's Hospital, Harvard Medical School, USA

Abstract. In this paper, we use Spherical Topic Models to discover the latent structure of lung disease. This method can be widely employed when a measurement for each subject is provided as a normalized histogram of relevant features. In this paper, the resulting descriptors are used as phenotypes to identify genetic markers associated with the Chronic Obstructive Pulmonary Disease (COPD). Features extracted from images capture the heterogeneity of the disease and therefore promise to improve detection of relevant genetic variants in Genome Wide Association Studies (GWAS). Our generative model is based on normalized histograms of image intensity of each subject and it can be readily extended to other forms of features as long as they are provided as normalized histograms. The resulting algorithm represents the intensity distribution as a combination of meaningful latent factors and mixing coefficients that can be used for genetic association analysis. This approach is motivated by a clinical hypothesis that COPD symptoms are caused by multiple coexisting disease processes. Our experiments show that the new features enhance the previously detected signal on chromosome 15 with respect to standard respiratory and imaging measurements.

1 Introduction

In this paper, we employ the Spherical Topic Model[1] (which is one of the variants of the latent topic models) to extract imaging features for genetic association studies. It is common in classical Genome-Wide Association Studies (GWAS) to perform statistical association between genetic measurements and a few quantities such as diagnosis. Imaging features provide rich information about the disease phenotype and promise to enhance the sensitivity of the genetic studies. Using individual voxels as a phenotype is not informative and due to the noisy nature of imaging measurements induces high false positive rate. Therefore, summarizing imaging features into meaningful quantities (*i.e.,* dimensionality reduction) improves the association and facilitate interpretation of the results. In this work, we build on a variant of topic models to perform this step of dimensionality reduction.

COPD is characterized by chronic and progressive difficulty in breathing, and is one of the leading causes of death in the United States [2]. The disorder is

M.J. Cardoso et al. (Eds.): BAMBI 2014, LNCS 8677, pp. 107–117, 2014.

believed to be a mixture of multiple disease processes including the destruction of the air sacs (emphysema) and inflammation of the airways (airway disease). Each process consists of multiple subtypes [3]. In this paper, we focus on emphysema which manifests itself as changes in intensity of the lung in Computed Tomography (CT) images [3]. Therefore, we use image intensity of the lung as a unit of measurements for each subject. The goal is to summarize this information into meaningful features. Similar to the idea of *bag of words* in natural language processing, later also adopted in computer vision [4], we view a histograms as a *document* and subtypes of the disease as different *topics*. This approach assumes that every patient (document) contains multiple portions of the disease subtypes (topics) and those disease subtypes, *i.e.,* topics, are shared across subjects. The goal of this paper is not to diagnose COPD since a test of lung function via forced exhalation has been the gold standard of COPD diagnosis for decades [5]. Our aim is to use imaging features to characterize the phenotype and the underlying genetic causes of the disease.

The search for genetic variants that increase the risk of a disorder is one of the central challenges in medical research, and has been traditionally performed via GWAS. Standard GWAS identifies correlations between genetic variants and a single phenotype (*e.g.,* mostly disease vs. control). Although such analysis identified several variants relevant to COPD (*e.g.,* IREB2 on chromosome 15 [6]), such studies are likely incomplete. First, COPD is a mixture of diseases and therefore is unlikely to be explained by a single factor. Second, the effect of the genetic variants may be scattered across the lung volume but their cumulative effect is manifested in the respiratory signal [7]. Imaging can help to address both challenges. Image features that capture the amount of emphysema have been previously demonstrated to reflect disease pathology and predict outcomes in COPD [7]. We seek to extract features from images that capture heterogeneous manifestations of the disease and enrich detection of genetic markers associated with COPD.

The standard approach to quantify emphysema is to apply an intensity threshold within the volume of the lung to compute a surrogate measure for the volume of emphysema [7]. Clinical studies suggest that lungs of COPD patients present symptoms of different subtypes of emphysema [7, 5]. Recent work exploits spatial patterns of intensity to classify emphysema into subtypes. Examples include the use of Kernel density estimation [8], combination of Local Binary Pattern (LBP) and intensity histogram [9], and Multi-coordinate Histogram of Oriented Gradient (MHOG) descriptors [10] for subtype classification of image patches in CT. Importantly, none of the method above characterizes how the underlying biological processes overlap with radiologic categorization.

Imaging genetics associates image phenotype with genetic markers relevant for the disease of interest. The objective is to characterize clinical heterogeneity of the disease and to detect novel genetic markers associated with COPD [11]. Most methodological innovations in imaging genetics to date have been demonstrated in the context of neuro-degenerative diseases [12, 13, 14], where image features are typically computed in a common coordinate system and are assumed to be spatially consistent across subjects. Unfortunately, such coordinate system does not exist for the lung, presenting an additional challenge for creating image-based descriptors that can be compared across subjects.

In this paper, we build a generative model that encodes the clinical assumption that COPD symptoms are caused by multiple coexisting biological processes. We assume that every subject is a mixture of latent disease factors, that are shared across the population. This approach is referred to as topic modeling in machine learning (*e.g.,* LDA [15]). The contribution of each latent factor for a particular subject becomes a new feature that can be used as an intermediate phenotype for detecting genetic associations. To integrate the resulting features into genetic analysis, we employ a method that views the genotype as the dependent variable and uses all the latent features simultaneously to find the genetic association. We demonstrate that the new features enhance the signal on chromosome 15 by improving the sensitivity of detection.

2 Topic Modeling for Feature Extraction

Previous studies have shown the intensity of lung to be highly informative for characterization of COPD [8, 9]. Therefore, we use global histogram of image intensity of the lung as a unit of measurement for each subject. The goal is to reduce a set of histograms to a set of meaningful features that enhance subsequent statistical analysis. Histogram data can in general encode richer features such as sophisticated localized descriptors (*e.g.,* Histogram of Oriented Gradients (HOG)), but to focus on the model, we limit ourselves to image histograms which have been shown to be informative for COPD [8, 9]. Here, we adopt the Spherical Admixture Model [1] that views each histogram as a point on a hypersphere. The advantage of this model is that it can handle unit-less (normalized) representations of the histograms. This property allows us to normalize the features by the volume of the lung.

We assume an image of subject n in a study is represented by a distribution $\mathbf{y}_n \in \mathbb{R}^D$ $(\sum_{d=1}^{D} y_{nd} = 1)$. With a change of the variables $y_{nd} := z_{nd}^2$, we map the intensity distribution to a unit hypersphere, $\mathbf{z}_n \in \mathbb{S}^{D-1}$. Motivated by the clinical hypothesis that COPD is a mixture of diseases, we assume that each data point (subject) is a normalized sum of K disease factors $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \cdots \boldsymbol{\phi}_K] \in \mathbb{R}^{D\times K}$. The factors are shared across the population and each factor is also a distribution, $\boldsymbol{\phi}_k \in \mathbb{S}^{D-1}$ $(1 \le k \le K)$. The generative model can be summarized as follows[1]:

$$\begin{aligned}
\boldsymbol{\mu} &\sim \mathrm{vMF}(\mathbf{m}, \kappa_0), \\
\boldsymbol{\phi}_k &\sim \mathrm{vMF}(\boldsymbol{\mu}, \xi), \\
\mathbf{x}_n &\sim \mathrm{Dirichlet}(\boldsymbol{\alpha}), \\
\mathbf{z}_n &\sim \mathrm{vMF}\left(\frac{\boldsymbol{\Phi}\mathbf{x}_n}{\|\boldsymbol{\Phi}\mathbf{x}_n\|_2}, \kappa \right)
\end{aligned} \tag{1}$$

where vMF(·) and Dirichlet(·) denote the von Mises-Fisher (vMF) [16] and Dirichlet distributions respectively. vMF distribution is a natural distribution, akin to a multivariate Normal distribution, for directions on a sphere. $\boldsymbol{\mu}$ is a latent variable that controls the mean of the disease factors (topics), $\mathbf{m}$ and κ_0 are hyper-parameters that define the mean and concentration of $\boldsymbol{\mu}$ respectively. $\mathbf{x}_n$ is a normalized latent distribution that defines a portion of each disease factor (topic) represented in subject n. Since $\mathbf{x}_n$ is normalized (sums to one), Dirichlet distribution is a reasonable prior choice; $\boldsymbol{\alpha}$ is the multivariate shape parameter of the Dirichlet distribution. $\frac{\boldsymbol{\Phi}\mathbf{x}_n}{\|\boldsymbol{\Phi}\mathbf{x}_n\|_2}$ maps the weighted sum of the topics back to the sphere and serves as a noiseless representation of the observation $\mathbf{z}_n$. To accommodate possible noise, the observation is modeled as a von Mises-Fisher perturbation of the noiseless representation, parameter κ controls the concentration of the noise. For notational convenience, we define $\Omega = \{\boldsymbol{\mu}, \boldsymbol{\Phi}, \mathbf{X}\}$ to be the set of the latent variables and $\Upsilon = \{\alpha, \xi, \kappa, \kappa_0\}$ to represent the set of hyper-parameters. The generative model is illustrated in Fig.1a.

The join probability $p(\mathbf{Z}, \Omega; \Upsilon)$ can be written as follows:

$$p(\mathbf{Z}, \Omega; \Upsilon) = \prod_{n=1}^{N} p(\mathbf{z}_n | \boldsymbol{\Phi}, \mathbf{x}_n; \Upsilon) p(\mathbf{x}_n; \Upsilon) \prod_{k=1}^{K} p(\boldsymbol{\phi}_k | \boldsymbol{\mu}; \Upsilon) p(\boldsymbol{\mu}; \Upsilon) \tag{2}$$

Reisinger *et al.* [1] proposed to use variational mean-field method to approximate the posterior distribution of the latent variables in this model with a fully factorized function as follows:

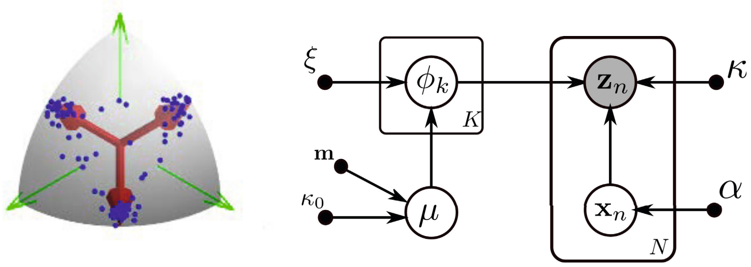
$$q(\boldsymbol{\mu}, \boldsymbol{\Phi}, \mathbf{X} | \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Phi}}, \tilde{\mathbf{m}}; \Upsilon) = q(\boldsymbol{\Phi} | \tilde{\boldsymbol{\mu}}, \xi) q(\mathbf{X} | \tilde{\boldsymbol{\alpha}}) q(\boldsymbol{\mu} | \tilde{\mathbf{m}}, \kappa_0), \tag{3}$$

where $\Sigma = \{\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Phi}}, \tilde{\mathbf{m}}\}$ are the parameters of the approximate posterior distribution $q(\cdot)$. Note that Σ and Ω are not identical since the former is the set containing parameters of the approximate posterior distribution while the latter is the set of latent variables in the original model. The variational method minimizes the KL-divergence between the approximating distribution and the join probability distribution to find the optimal setting of the parameters:

$$(\Sigma^*, \Upsilon^*) = \arg\min_{\Sigma, \Upsilon} \mathbb{E}_q \left[\log p(\mathbf{Z}, \Omega; \Upsilon) - \log q(\Omega; \Upsilon) \right]. \tag{4}$$

Computing the derivatives with respect to Σ and Υ and setting them to zero, the mean field method reduces to a set of fixed-point update equations (see [1] for detail).

We seek to estimate the posterior means of the latent features $\hat{\mathbf{x}}_n := \mathbb{E}_q[\mathbf{x}_n]$, which serve as a low-dimensional representation of subject n, and are used to infer associated genetic markers of the disease as described in Section 3. Estimates $\hat{\mathbf{x}}_n$ can be viewed as a K-dimensional histogram defined over K latent factors. Indeed, we reduce the original D-dimensional histograms of image intensities to the K-dimensional histograms of the latent factors. Other quantities of interest are the latent factors, $\tilde{\boldsymbol{\phi}}_k$, which are D-dimensional histograms that describe each latent factor in the intensity space. The hyper-parameters, Υ, and the parameters of the approximate posterior distribution, Σ, are estimated during learning (*i.e.,* Eq. 4). The main parameter of the method is number of topics K.



(a) (b)

Fig. 1. (a) Schematic visualization of the generative model. Each data point (blue) is a noisy mixture of latent disease factors (red arrows). (b) Graphical model for the spherical topic model in [1]. The open gray and white circles are the observed and the latent random variables respectively. The full circles are the hyper-parameters.

Unlike traditional Factor Analysis methods such as PCA, this approach yields normalized factors and coefficients (*i.e.,* both can be interpreted as histograms). This is advantageous for interpretation of the results because the $\mathbf{\Phi}$ can be viewed the same way as the input histograms and mixing weights $\mathbf{x}_n$ can be viewed as the proportions of each factor in subject n.

3 From Image Features to Genetic Markers

In addition to the image features $\hat{\mathbf{x}}_n$, each subject is characterized by a vector of S genetic markers ($g_{ns} \in \{0, 1, 2\}$, $1 \le s \le S$). g_{ns} represents the allele count in the locus s of the genetic measurement for subject n. Standard GWAS builds a regression model $\hat{x}_{nk} = b_{s,k} + w_{sk} g_{ns} + \varepsilon_{nsk}$ for each Single Nucleotide Polymorphism (SNP) g_{ns} and the phenotype $\hat{x}_{nk}$ separately. The detection procedure aims to reject the null hypothesis of no association ($w_{sk} = 0$) by performing t-test. Contrary to the standard GWAS that models phenotype as a dependent variable, we use a previously proposed method that considers the genotype as the dependent variable and uses all phenotypes features simultaneously [17]. The algorithm employs proportional odds (ordinal) logistic regression to model the allele count. Unlike multi-class logistic regression, ordinal logistic regression assumes the classes (*i.e.,* $g_{ns} = 0, 1, 2$) are ordered, the hyperplanes separating the classes are parallel, and the difference between classes is captured by the intercepts as illustrated in Fig.2b,2a. Ordinal logistic regression is more restrictive than a more general multi-class logistic regression and exhibits fewer degrees of freedom. The ordinal method is more appropriate when a natural ordering can be imposed on class labels. This is certainly the case here since g_{ns} counts the number of minor alleles and we assume an additive effect. The cumulative probability is modeled as the logistic function:

$$\mathbb{P}(g_{ns} \le j) = \psi(\mathbf{w}_s^T \hat{\mathbf{x}}_n - b_{s,j}) = \frac{1}{1 + \exp(\mathbf{w}_s^T \hat{\mathbf{x}}_n - b_{s,j})}, \tag{5}$$

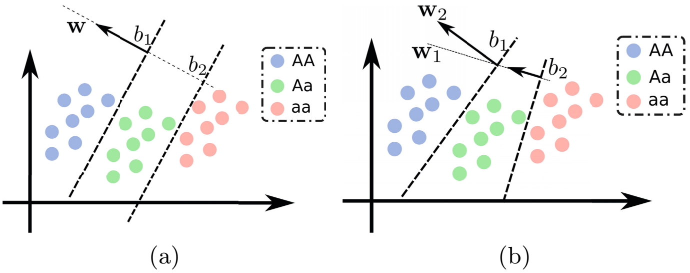


Fig. 2. (a) Ordinal vs. (b) Multi-class logistic regression. In the ordinal regression, the separating hyperplanes are parallel (same $\mathbf{w}$) and classes differ by intercepts.

where $j \in \{0, 1, 2\}$. For the allele j in locus s, we estimate one weight $\mathbf{w}_s$ and two intercepts $b_{s,1}$ and $b_{s,2}$. Fitting the model reduces to maximizing the log-likelihood of data to find the best parameters ($\mathbf{w}_s$, $b_{s,1}$, $b_{s,2}$) for each SNP g_{ns}:

$$\mathcal{L}(\mathbf{w}_s, b_{s,1}, b_{s,2}; \hat{\mathbf{x}}) = \sum_{n=1}^{N} \log \left(\psi(\mathbf{w}_s^T \hat{\mathbf{x}}_n - b_{s,(g_n+1)}) - \psi(\mathbf{w}_s^T \hat{\mathbf{x}}_n - b_{s,g_n}) \right), \quad (6)$$

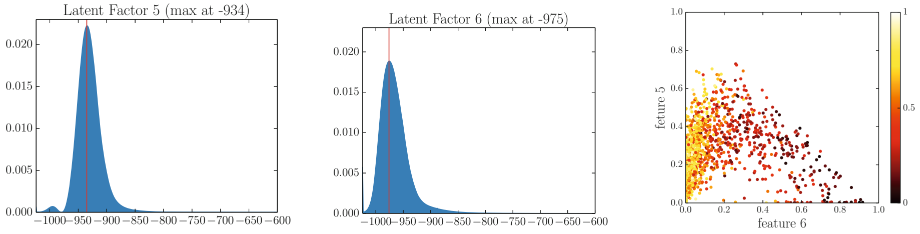
where $b_{s,0} = -\infty$ and $b_{s,3} = +\infty$.

We compute the likelihood ratio of the model with combination of covariates and $\hat{\mathbf{x}}$ ($\mathcal{H}_1$) versus only the covariates ($\mathcal{H}_0$). χ^2 distribution with degrees of freedom equal to the difference in dimensionality is used to compute the p-value [17]. Covariates are defined in the next section.

4 Experiments

Experiments in this section are organized as follows. We first qualitatively evaluate the new features $\hat{\mathbf{x}}_n$ and the estimated latent factors $\tilde{\boldsymbol{\phi}}_k$ (Fig.3). Next, we select a few important SNPs to study the sensitivity of the algorithm with respect to the model size K (Fig.4). Finally, we study how much the new features enrich our genetic findings versus the traditional measurements such as airflow (Fig.5 and Fig.6).

Data: We demonstrate the method on a large COPD study of 6,670 subjects. The respiratory measurements include: percent predicted, forced expiratory volume in one second (FEV_1) that is used as an indicator of COPD severity, and the ratio of FEV_1 over forced vital capacity (FEV_1/FVC), used as a measure of airflow obstruction for COPD diagnosis. We will refer to the respiratory measures as `Resp`. We also evaluate summary measurements computed from lung CT. These include percent emphysema, defined as the percentage of lung tissue below -950 Hounsfield units; percent gas trapping, defined as the percentage of lung tissue below -910 Hounsfield units after exhalation, and the wall thickness of an airway with an internal perimeter of $10mm$ (Pi_{10}). We will refer to



(a) (b) (c)

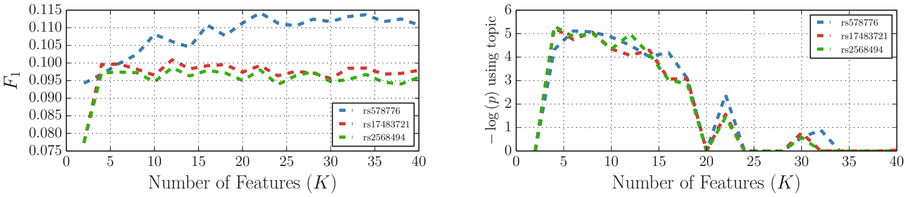
Fig. 3. Estimated latent model estimated. (a),(b) Examples of latent factors for $K = 6$. (c). Scatter plot of latent features colored by FEV_1/FVC (severity of COPD). Hotter colors represent higher values (healthier subjects). The scatter plots show that new features successfully delineate the severity of the disease.

these measures as `sumImg`. The subjects were genotyped by Illumina on the HumanOmniExpress array. We employ standard quality control for genetic data, including missing-ness, excess heterozygous, gender mismatch, cryptic relatedness, population outliers, marker concordance, and Hardy-Weinberg equilibrium. We computed 6 principal components from the genotype to correct for population heterogeneity, and included them in the covariate set along with age, Body Mass Index (BMI) and number of aggregate packs smoked per year.

Qualitative Evaluation: Fig. 3 shows examples of the derived latent disease factors ($\tilde{\boldsymbol{\phi}}_k$) and the corresponding latent features ($\hat{x}$) in the patient cohort. As shown in Fig.3a and Fig.3b, every factor is a proper distribution. In effect, the classical method is based on a single threshold that divides a histogram into two bins: lower or higher bins. There is a debate in the COPD community on what the optimal threshold should be. In contrast to the traditional approach, one can view the proposed method as an adaptive way of histogram binning with no need to specify the threshold explicitly. Nevertheless, it is interesting to see that the latent factors are located at the values that are close to -950 Hounsfield units (-950 is commonly used to define percentage of emphysema in the COPD community).

Fig.3c presents a scatter plot of pairs of new features ($\hat{x}$) in the cohort. The color in the scatter plot indicates the value of FEV_1/FVC. Higher values correspond to subjects without COPD. The scatter plot suggests that the new features successfully characterize the severity of the disease. Notice the smooth variation across the population. We also performed linear regression between new features ($K = 6$) and respiratory measurement FEV_1 ($R^2 = 0.67$), FEV_1/FVC ($R^2 = 0.74$), and the percent of emphysema ($R^2 = 0.96$).

Sensitivity Analysis: We chose around 500 SNPs with the lowest p-values identified in previous studies. Many of these SNPs are from regions that have been frequently reported in the genetic and respiratory literature in connection



(a)

(b)

Fig. 4. Prediction accuracy (F_1-measure) for 5-fold cross validation and quality-of-fit ($-\log(p)$) as a function of the model size K for three important genetic markers. F_1 is defined as 2(precision × recall)/(precision + recall). We note the improvements in prediction accuracy in (a). As the model becomes more complex (higher K), the number of degrees of freedom in χ^2 distribution increases, which explains the initial increases and decreases in the p-value in (b).

to lung cancer genes or nicotine receptors areas. We first examine the behavior of the algorithm on the smaller set of 2,441 subjects. In order to study the sensitivity of the method with respect to the main parameter (the number of the latent factors K), we choose three SNPs associated with COPD (`rs578776` [18]), nicotine dependence (`rs17483721` [19]), and lung cancer (`rs2568494` [6]), and evaluate the significance of the model fit for different values of K.

The cross-validation accuracy of the model saturates very fast (Fig.4a) implying that few topics summarize the dataset successfully. As K grows, so does the number of degrees of freedom in the χ^2 distribution that is used to evaluate the significance of the fit in Fig.4b. Unless the fit improves substantially, we expect the significance ($-\log(p)$) to increase at first and then to decline. The plots in Fig.4b spike down at $K = 20, 24, 34$ because the features become so collinear that the optimization of the cost function of the ordinal logistic in Eq. (5) does not converge (Hessian in Eq. (6) become ill-conditioned). An alternative way to choose K is to use the variational lower bound which is not explored in this paper.

Association Study: To test if the new features enrich the association, we examined different combinations of topic features, summary image features (`sumImg`) and the respirometry measurement (`Resp`) for the set of selected SNPs. Fig. 5 reports the pair-wise comparison of different feature sets. $A > B$ indicates how many more SNPs are detected in one setting (A) versus the other (B) and how they were distributed across different chromosomes. Almost every combination with `latFtr` improves with respect to the second row (`sumImg`). We conclude that the extracted features are correlated with previously identified clinical image-based measures, but also offer complementary detections for genetic studies. Another important message from Fig. 5 is that adding the most important clinical measurement (`Resp`) improves the results.

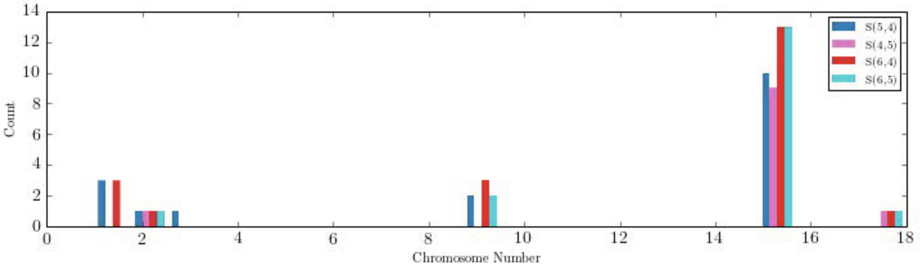


Fig. 5. Comparison of different feature sets. For $K = 4$ and for different combination of features, $A > B$ indicates how many more SNPs are detected in one setting (A) versus the other (B) and how they were distributed in the different chromosomes.

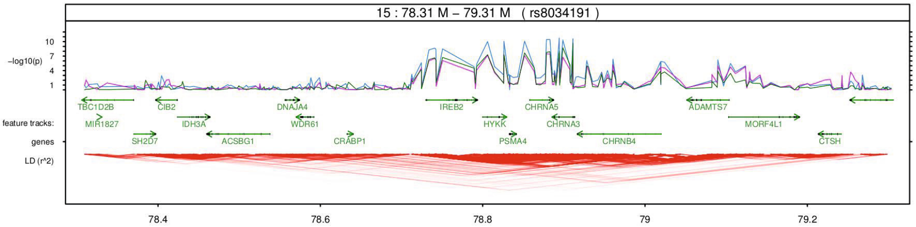


Fig. 6. Fine-scale regional maps for the region of significance on chromosome 15. Blue, purple and green lines represent `latFtr`, `sumImg`, and `Resp` respectively.

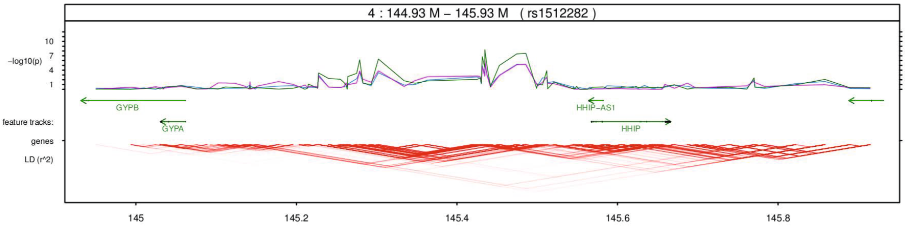


Fig. 7. Fine-scale regional maps for the region of significance on chromosome 4. Blue, purple and green lines represent `latFtr`, `sumImg`, and `Resp` respectively. There is signal that is only detected effectively by the respiratory features but not by `sumImg` or `latFtr`.

We also extracted features for the whole set of 6,670 subjects and applied regression on the genome-wide scale. Fig.6 shows the regional maps on the chromosomes 15. Blue, purple and green lines represent new features (`latFtr`), `sumImg`, and `Resp` features. On the chromosomes 15, the new features (`latFtr`) enhanced the detection with respect to the other two feature sets by about 4 orders of magnitude in the corresponding p-values. On the chromosome 4, there is signal that is only detected effectively by the respiratory features but not by `sumImg` or latFtr (see Fig.7). This suggests there is some information in the respiratory signal that is not reflected in the images.

5 Conclusion

Traditional approaches to CT analysis in lung disease often rely on a single threshold or set of thresholds, and ignore the effects of genetic variants. We present a method to extract image features using topic modeling from lung CT images. Bins of the histogram are viewed as words in a dictionary or codebook. Our experiments show that new features correlate well with clinical measures of physiology (spirometry) and generalize commonly used measures for emphysema. The new features promise to improve the power of genetic associations for genetic causes of COPD. The proposed method is general and can be applied to any distribution. Including texture and lobe information to better characterize different subtypes of emphysema is a clear important and promising direction of future research.

Acknowledgements. This work was supported by NIH NIBIB NAMIC U54-EB005149, NIH NCRR NAC P41-RR13218 and NIH NIBIB NAC P41-EB-015902, NHLBI R01HL089856, R01HL089897, K08HL097029 and R01HL113264. The COPDGene study (NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, Siemens and Sunovion.

References

[1] Reisinger, J., et al.: Spherical Topic Models. In: Fürnkranz, J., Joachims, T. (eds.) ICML, pp. 903–910. Omnipress (2010)

[2] Regan, E.A., et al.: Genetic epidemiology of COPD (COPDGene) study design. COPD: Journal of Chronic Obstructive Pulmonary Disease 7(1), 32–43 (2011)

[3] Satoh, K., Kobayashi, T., Misao, T., Hitani, Y., Yamamoto, Y., Nishiyama, Y., Ohkawa, M.: Ct assessment of subtypes of pulmonary emphysema in smokers. CHEST Journal 120(3), 725–729 (2001)

[4] Sivic, J., Zisserman, A.: Efficient visual search of videos cast as text retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(4), 591–606 (2009)

[5] Thurlbeck, W.M., et al.: Emphysema: definition, imaging, and quantification. AJR. American Journal of Roentgenology 163(5), 1017–1025 (1994)

[6] Guo, Y., et al.: Genetic analysis of IREB2, FAM13A and XRCC5 variants in Chinese Han patients with chronic obstructive pulmonary disease. Biochemical and Biophysical Research Communications 415(2), 284–287 (2011)

[7] Castaldi, P.J., et al.: Distinct quantitative computed tomography emphysema patterns are associated with physiology and function in smokers. American Journal of Respiratory and Critical Care Medicine 188(9), 1083–1090 (2013)

[8] Mendoza, C.S., et al.: Emphysema quantification in a multi-scanner HRCT cohort using local intensity distributions. In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 474–477. IEEE (2012)

[9] Sorensen, L., et al.: Lauge: Quantitative analysis of pulmonary emphysema using local binary patterns. IEEE Transactions on Medical Imaging 29(2), 559–569 (2010)

[10] Song, Y., et al.: Feature-Based Image Patch Approximation for Lung Tissue Classification. IEEE Trans. Med. Imaging 32(4), 797–808 (2013)

[11] Manichaikul, A., et al.: Genome-wide Study of Percent Emphysema on CT in the General Population: The MESA Lung/SHARe Study. American Journal of Respiratory and Critical Care Medicine (ja) (2014)

[12] Batmanghelich, N.K., Dalca, A.V., Sabuncu, M.R., Golland, P.: Joint modeling of imaging and genetics. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 766–777. Springer, Heidelberg (2013)

[13] Filippini, N., et al.: Anatomically-distinct genetic associations of APOE e4 allele load with regional cortical atrophy in Alzheimer's disease. Neuroimage 44(3), 724–728 (2009)

[14] Vounou, M., et al.: Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. NeuroImage 53(3), 1147–1159 (2010)

[15] Blei, et al.: Latent dirichlet allocation. The Journal of machine Learning research 3, 993–1022 (2003)

[16] Dhillon, I.S., et al.: Modeling Data using Directional Distributions. Technical Report TR-03-06, The University of Texas at Austin (January 2003)

[17] O'Reilly, P.F., et al.: MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PLoS ONE 7(5), e34861 (2012)

[18] Saccone, N.L., et al.: Multiple independent loci at chromosome 15q25. 1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. PLoS genetics 6(8), e1001053 (2010)

[19] Hung, R.J., et al.: A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 452(7187), 633–637 (2008)

A Generative Model for Automatic Detection of Resolving Multiple Sclerosis Lesions

Colm Elliott[1], Douglas L. Arnold[3], D. Louis Collins[2], and Tal Arbel[1]

[1] Centre for Intelligent Machines, McGill University, Canada
[2] Montreal Neurological Institute, McGill University, Canada
[3] NeuroRx Research, Montreal, Canada

Abstract. The appearance of new Multiple Sclerosis (MS) lesions on MRI is usually followed by subsequent partial *resolution*, where portions of the newly formed lesion return to isointensity. This resolution is thought to be due mostly to reabsorption of edema, but may also reflect other reparatory processes such as remyelination. Automatic identification of resolving portions of new lesions can provide a marker of repair, allow for automated analysis of MS lesion dynamics, and, when coupled with a method for detection of new MS lesions, provide a tool for precisely measuring lesion change in serial MRI. We present a method for automatic detection of resolving MS lesion voxels in serial MRI using a Bayesian framework that incorporates models for MRI intensities, MRI intensity differences across scans, lesion size, relative position of voxels within a lesion, and time since lesion onset. We couple our method with an existing method for automatic detection of new MS lesions to provide an automated framework for measuring lesion change across serial scans of the same subject. We validate our framework by comparing to lesion volume change measurements derived from expert semi-manual lesion segmentations on clinical trial data consisting of 292 scans from 73 (54 treated, 19 untreated) subjects. Our automated framework shows a) a large improvement in segmentation consistency over time and b) an increased effect size as calculated from measured change in lesion volume for treated and untreated subjects.

1 Introduction

The appearance of new Multiple Sclerosis (MS) lesions visible on T2-weighted MRI is generally followed by a period of repair or lesion *resolution*, during which portions of the new lesion will return towards isointensity on MRI [1]. This resolution is thought to be due mostly to reabsorption of edema, but may also reflect other reparatory processes such as remyelination [1]. The percentage of new lesion that resolves has been posited as a marker for tissue repair and for staging disease [1]. Meier et al. have previously modeled the dynamics of new lesion formation on T2-weighted MRI and have observed a transient phase of 3-4 months, with larger lesions exhibiting a proportional greater amount of lesion resolution, and concentric patterns of resolution where voxels near the lesion boundaries are much more likely to resolve than those in the lesion center [1,2].

M.J. Cardoso et al. (Eds.): BAMBI 2014, LNCS 8677, pp. 118–129, 2014.

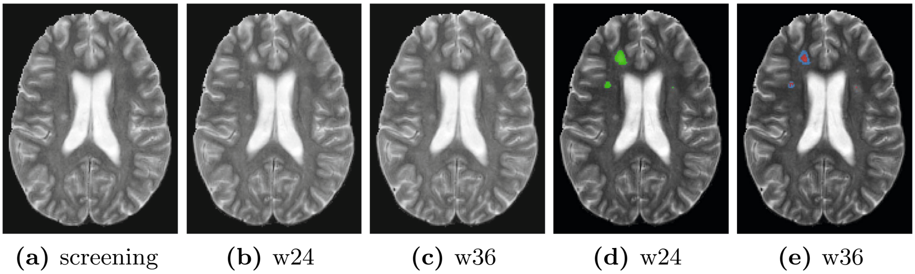


Fig. 1. Example of new and resolving MS lesion on T2-weighted MRI over 3 timepoints. (a)-(c) show T2w images at screening, week 24 and week 36, while (d) and (e) overlay new and resolving lesion voxels at w24 and w36, where green denotes new, red denotes stable portions of new lesion and blue denotes resolving portions of new lesion, with respect to the previous timepoint.

An example of lesion formation and resolution over 3 serial scans is shown in Fig. 1.

Manual segmentation of MS lesion on MRI is time-consuming and subject to inter and intra-rater variability. Although many methods for automatic segmentation have been proposed [3], they remain imperfect, generally require substantive manual correction in real-world clinical environments, and still have relatively high degree of variability. Additionally, most methods do not take advantage of temporal relationships when considering multiple timepoints of the same subject, leading to reduced sensitivity to change and higher temporal segmentation variability, thus confounding inconsistent segmentation with real biological change. Several approaches have been proposed for the automatic segmentation of *new* MS lesions in sequential MRI [4], but little has been proposed for automatic detection of lesion resolution. While lesion resolution is implicitly modeled in [5], spatial and temporal characteristics of the resolution process are not modeled and the validation focuses exclusively on the detection of new lesions.

In this paper, we present a novel method for automatic detection of resolving lesion. A generative Bayesian model is used to detect resolving portions of lesions, where we consider MRI intensities, MRI intensity differences across time (difference images) and where we embed previously observed characteristics of lesions formation such as lesion size, time from lesion onset, and relative positions of voxels within a lesion [1].

Meaningful validation of any lesion segmentation algorithms is difficult due to the absence of a real ground truth. While manual references are often used for comparisons [3], these are generally imperfect, highly variable, and time-consuming to generate. The variability of lesion segmentations over timepoints of the same subject also makes them impractical as a basis for generating a reference for resolving lesion as most of the apparent resolution from one timepoint to the next would be attributable to inconsistent lesion boundaries rather than to veritable biological change. In the absence of an explicit reference segmentation for resolving lesion voxels, we have chosen to combine our method with a method for new lesion detection [6] to provide a method for detection of lesion *change* in serial MRI. We compare lesion change measurements generated from our method to those derived from semi-manual reference segmentations lesions generated independently at each timepoint. We validate our method by comparing a) segmentation consistency across time and b) apparent treatment benefit as determined by effect size calculated from lesion volume change measurements from treated (N=54) and untreated (N=19) subjects in our test data.

2 Method

2.1 Bayesian Formulation

We present a Bayesian framework for automatic detection of resolving portions of lesions in serial MRI. We use a generative model where, at each voxel i in a lesion, we consider MRI intensities, $\boldsymbol{I_i^t}$, at the current timepoint and intensity differences, $\boldsymbol{D_i^t}$, between coregistered current and previous timepoints. We additionally consider the distance from lesion boundary, d_i, to model a concentric pattern of resolution, and lesion size at onset, s, to model the increased relative rates of resolution of larger new lesions. Finally we consider the time from lesion onset, a, to model the fact that most resolution occurs soon after lesion onset [7]. We define lesion onset as the time of first observation of a new lesion.

Resolution of Recently New Lesion

We first consider the case where we are provided with a set of new MS lesions that appear after our first available timepoint for a given subject, such that we can determine time of onset. In practice, these recently new lesions will be generated by an automated method as in [6]. We consider each lesion in the set of recently new lesions in turn, inferring the probability of resolution at each voxel i of the lesion, at all timepoints following lesion onset. Lesion size and boundaries are determined at lesion onset.

We allow two states for resolution status, res_i^t: a) *resolved*, corresponding to lesion that returned to "healthy" tissue from lesion at time t ($res_i^t = 1$), and b) *stable*, lesion which remains lesion at time t ($res_i^t = 0$). The distance from lesion boundary, d_i, is normalized based on lesion size and takes on a value between 0 (closest to lesion edge) and 1 (furthest from lesion edge) to provide invariance to lesion size.

For each voxel i in a given lesion, we wish to determine the probability of resolution at time t, based on observed MRI intensities at time t, $\boldsymbol{I_i^t}$, MRI intensity differences between times t and $t-1$, $\boldsymbol{D_i^t}$, as well as time since lesion onset, a, lesion size at onset, s, and distance from the lesion boundary, d_i:

$$
\begin{aligned}
&p(res_i^t|\boldsymbol{I_i^t}, \boldsymbol{D_i^t}, s, d_i, a) = \\
&= \frac{p(\boldsymbol{I_i^t}|res_i^t, \boldsymbol{D_i^t}, s, d_i, a)p(\boldsymbol{D_i^t}|res_i^t, s, d_i, a)p(s|res_i^t, d_i, a)p(d_i|res_i^t, a)p(res_i^t|a)}{p(\boldsymbol{I_i^t}|\boldsymbol{D_i^t}, s, d_i, a)p(\boldsymbol{D_i^t}|s, d_i, a)p(s|d_i, a)p(d_i|a)} \\
&= \frac{1}{K} p(\boldsymbol{I_i^t}|res_i^t)p(\boldsymbol{D_i^t}|res_i^t)p(s|res_i^t)p(d_i|res_i^t)p(res_i^t|a). \qquad (1)
\end{aligned}
$$

We have invoked Bayes' law multiple times, have treated the denominator as a normalization constant, and have made several statistical conditional independence assumptions:

- MRI intensity at time t is conditionally independent of $\boldsymbol{D_i^t}$,a,s, and d_i, given resolution status $(p(\boldsymbol{I_i^t}|res_i^t, \boldsymbol{D_i^t}, s, d_i, a)) = p(\boldsymbol{I_i^t}|res_i^t))$.
- MRI intensity difference $\boldsymbol{D_i^t}$ is conditionally independent of a,s, and d_i, given resolution status $(p(\boldsymbol{D_i^t}|res_i^t, s, d_i, a)) = p(\boldsymbol{D_i^t}|res_i^t))$.
- The lesion size is independent of normalized distance from lesion boundary and time from onset, given resolution status $(p(s|res_i^t, d_i, a) = p(s|res_i^t))$.
- The normalized distance from lesion boundary is conditionally independent of time from onset, given resolution status $(p(d_i|res_i^t, a) = p(d_i|res_i^t))$.

Our posterior probability of resolution at voxel i at time t is thus a product of 5 terms, each of which models the likelihood of resolution status based on one of intensity, intensity difference, distance from lesion boundary, lesion size, and time since lesion onset.

Lesion Resolution with Limited Scan History

In some instances, we may have no or insufficient scan history to determine which set of existing lesions are new and which are not (e.g. at first timepoint). In such cases, we assume that we are given a segmentation of all lesions at the first timepoint and we attempt to jointly infer which lesions are *recently new* at the first timepoint and resolution status at subsequent timepoints. We use RN_i^t to denote whether voxel i at time t corresponds to lesion that is recently new ($<$ 6 months old) or not. Lesion newness, RN_i^t, is inferred by considering MRI intensities at the current timepoint and intensity differences between the current and ensuing timepoint, based on the observations that transient new lesions exhibit greater hyperintensity than stable older lesions and that this hyperintensity will decrease over time.

We can express the probability of being a recently new lesion at time $t-1$ based on MRI intensities at t and intensity difference between time t and time $t-1$ as:

$$p(RN_i^{t-1}|\boldsymbol{I_i^{t-1}}, \boldsymbol{D_i^t}) = \frac{1}{K} p(\boldsymbol{I_i^{t-1}}|RN_i^{t-1})p(\boldsymbol{D_i^t}|RN_i^{t-1})p(RN_i^{t-1}), \qquad (2)$$

where we have made conditional independence assumptions equivalent to those made for Eq. 1.

The boundaries of new lesions for which onset can be observed (i.e. at second or later timepoints) can be reliably determined even if they are confluent with existing lesion. However, for lesions where onset is not observed (i.e. lesions present at baseline), we cannot always reliably determine boundaries of individual lesions based on connectedness, especially for subjects with relatively high lesion load. As such, we choose not to consider lesion size and relative position of voxels in a lesion when determining probability of resolution in cases with insufficient scan history, as was done in Eq. 1. Incorporating our inference of lesion *newness*, we then infer resolution status for cases with insufficient scan history as:

$$\begin{aligned} p(res_i^t, RN_i^{t-1} | \boldsymbol{I_i^t}, \boldsymbol{D_i^t}, \boldsymbol{I_i^{t-1}}, a) &= p(res_i^t | RN_i^{t-1}, \boldsymbol{I_i^t}, \boldsymbol{D_i^t}, a) p(RN_i^{t-1} | \boldsymbol{I_i^{t-1}}, \boldsymbol{D_i^t}) \\ &= p(res_i^t | \boldsymbol{I_i^t}, \boldsymbol{D_i^t}, a) p(RN_i^{t-1} | \boldsymbol{I_i^{t-1}}, \boldsymbol{D_i^t}), \qquad (3) \end{aligned}$$

where the right side of Eq. 3 is our inference of newness at time $t-1$ as determined by Eq. 2, and the left side is our inference on resolution determined as in Eq. 1 but without using size and distance from lesion boundary. Here we assume a time since lesion onset equal to the difference between time t and time $t-1$ and also assume that only lesion voxels inferred as new at time $t-1$ are candidates for subsequent resolution.

3 Experiments

3.1 Data Sets

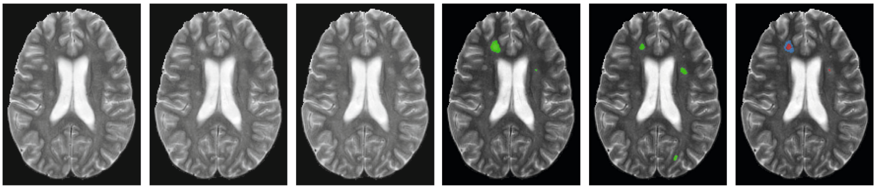
We use a proprietary clinical trial data set in our experiments, consisting of 639 scans from 73 subjects with relapsing-remitting MS, where each subject considered minimally had scans at screening (s), week 24 (w24), week 36 (w36), and week 48 (w48). Most subjects had additional intermediate scans at some or all of w04, w12, w16 and w20. T1-weighted with (T1c) and without (T1w) gadolinium injection, T2-weighted (T2w), proton-density weighted (PDw), and T2w Fluid-Attenuated Inversion Recovery (T2w-FLAIR) scans were available at each timepoint. All scans were acquired axially with an in-plane resolution of 1mm and slice thickness of 3mm, underwent non-uniformity correction [8], brain masking [9] and were rigidly registered across MRI modalities and timepoints [10]. Additionally, all scans underwent a decile-based piecewise linear intensity normalization to a global intensity space [11]. Semi-manual lesion segmentations of all MS lesions were performed at some timepoints by trained experts prior to and independently of this study, where an initial segmentation of MS lesion was generated using [12] and then manually corrected following a strict protocol. Semi-manual segmentations were available for all 73 patients but only for timepoints screening, w24, w36, and w48. As such only these 4 timepoints were used for validation (292 scans total). Any additional intermediate timepoints were included in the unsupervised training process. Treatment codes were made available for our data set, which identify subjects as either receiving treatment (N=54), or placebo (N=19). The availability of treatment codes allowed for validation based on effect size calculations, where we measure the statistical power of different measurements to differentiate treated and untreated (placebo) subjects.

Model Generation

Because we do not have ground truth for resolving lesion in our data set, we use an unsupervised approach to model learning. We first make use of existing software [6] to automatically identify new lesions in the entirety of our data set. This set of MS lesions, $\mathcal{L}_{new}$, then become candidates for resolution at timepoints following their appearance. We will consider model learning for inferring lesion resolution and inferring lesion *newness* separately, as different procedures are used for each.

Lesion Resolution Models
We use a hybrid unsupervised learning method where we first identify a set of representative samples of stable and resolving lesion voxels in our data, which we use to generate distributions for our intensity models, $p(\boldsymbol{I_i^t}|res_i^t)$, and intensity difference models, $p(\boldsymbol{D_i^t}|res_i^t)$. We then use these intensity based models to initialize our inference of resolution status at all voxels in $\mathcal{L}_{new}$ at timepoints following lesion onset, and use a generalized EM framework to learn parameters of our models for lesion size, normalized distance from lesion boundary, and lesion resolution conditioned on time since lesion onset.

Representative samples for stable and resolving lesions used for model initialization are generated using an approach based on difference of new lesions over 3 consecutive timepoints, as illustrated in Fig. 2. While such an approach is useful

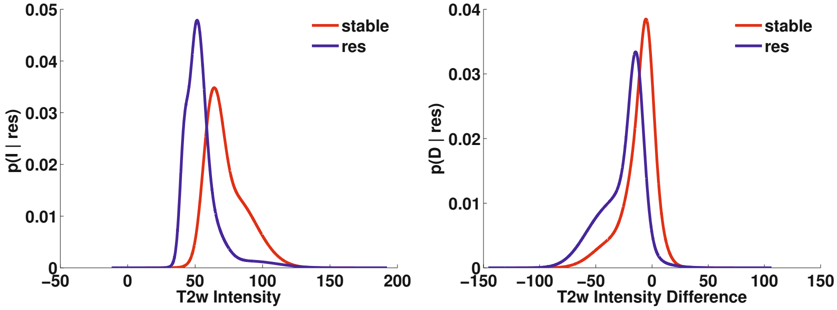


(a) $t1$ **(b)** $t2$ **(c)** $t3$ **(d)** new $t2$ **(e)** new $t3$ **(f)** stable / res

Fig. 2. Generating resolving and stable lesion samples using difference of new lesions. (a)-(c) show T2-weighted MRI of the same subject for 3 consecutive timepoints. (d)-(e) show new lesion (green) at $t2$ and $t3$ respectively, both with reference to $t1$. (f) shows resolving (blue) and stable (red) lesions samples at $t3$, where stable are those voxels that are identified as new in both (d) and (e), and resolving are those identified in (d) but not in (e).

for generating a set of samples, it is not a practical approach to detection of lesion resolution in the general case when considering additional (>3) timepoints, as it does not enforce temporal segmentation consistency and is not well suited to considering resolution occuring over multiple sequential timepoints.

Intensity based models are represented by 5 dimensional (corresponding to 5 modalities), 6-component Gaussian Mixture Models (GMMs), where the number of GMM components was chosen heuristically. Figure 3 shows learned probability densities for intensity and instensity differences, for resolving and stable (non-resolving) lesion voxels, marginalized over T2-weighted intensities.



(a) (b)

Fig. 3. (a) Intensity and (b) Intensity Difference densities for resolving and stable lesion voxels, shown marginalized over t2w for visulation purposes. In practice, models 5-dimensional densities corresponding to the 5 modalities used (T1w, T1c, T2w, PDw and T2w-FLAIR).

We use our intensity and intensity difference models to initialize resolution status of all voxels in $\mathcal{L}_{new}$ at timepoints following lesion onset. These initial estimates of resolution status will then act as hidden parameters in our EM learning framework. Model parameters for all our other models are then iteratively updated using generalized EM. The model learning process can be summarized as follows:

1. Generate a set of samples for resolving and stable lesion voxels using difference of new lesion.
2. Use samples generated in step 1 to generate models for intensity and intensity difference, for stable and resolved lesion voxels.
3. Generate a set of new MS lesions over all timepoints as candidates for resolution.
4. Initialize resolution status of new lesions generated in step 3 at all timepoints following onset, using only intensity and intensity difference models.
5. Initialize lesion size, normalized distance from lesion boundary, and resolution given time from onset models based on resolution status generated in step 4.

6. Use EM to iteratively update models, using models generated in step 5 as initializations. The E-step recalculates the probabilistic resolution status of new lesions based on the latest models and the M-step determines maximum likelihood models based on the updated resolution status.

We use a histogram based representation for our lesion size model, with 7 size ranges considered. We use exponential distributions to model both our normalized distance from boundary model and our resolution prior conditioned on lesion age.

Our EM framework jointly infers resolution status and model parameters. In our experiments, we have considered samples from the entirety of our data set (639 scans) for unsupervised model learning, but have only validated inference of resolution status in the subset of our data for which semi-manual lesion segmentations were available for comparison. Applying our learned models to a subset of the data from which they were learned can be considered as an additional E-step in our unsupervised learning framework.

Models for determining recently new lesion without scan history
We generate models for inferring lesion *newness* by generating intensity and intensity difference distributions for new and old (not recently new) lesions. We identify *old* lesions by considering voxels from a reference baseline lesion mask that remain lesion at least 6 months after baseline, or voxels from $\mathcal{L}_{new}$ that have not resolved 6 months after lesion onset. To identify *new* lesions we consider samples from $\mathcal{L}_{new}$ only at lesion onset and the subsequent timepoint. We again use 6 component GMMs to model our densities over the 5 modalities under consideration. Intensity and intensity difference models for new and old lesions are shown in Figure 4.

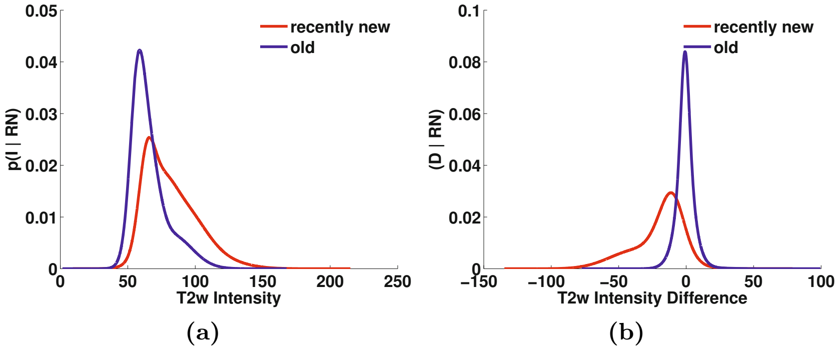


Fig. 4. (a) intensity and (b) intensity difference densities for new and old lesions, marginalized over t2w for visualization purposes. In practice, models 5-dimensional densities corresponding to the 5 modalities used (T1w, T1c, T2w, PDw and T2w-FLAIR).

3.2 Validation

Evaluating lesion resolution accuracy directly is not feasible due to the temporal variability in the available semi-manual reference lesion segmentations: most voxels identified as non-lesion at a given timepoint but as lesion in the previous timepoint would not actually correspond to resolving lesion but rather be due to temporal segmentation variability. As such, our validation focuses on measurements of lesion volume change over time, where we have coupled our method for detection of lesion resolution with an existing method for new lesion detection [6] to create a pipeline for lesion change detection in serial MRI based on change detection. We use the semi-manual reference segmentation as an initial lesion mask at our first timepoint, and lesion segmentation at subsequent timepoints is driven by detection of new, resolving and non-resolving (i.e. stable) lesion. Both lesion present at baseline and subsequently detected new lesions become candidates for resolution as determined by the proposed method. We compare the longitudinal segmentation of lesions as generated by our proposed pipeline based on change detection to the pre-existing semi-manual reference lesion segmentations. We validate based on a) segmentation consistency over time, and b) statistical power to differentiate treated from untreated subjects based on lesion volume change measurements.

Lesion Segmentation Consistency

Segmentation consistency is important as increased temporal variability will lead to less precise measurements of change over sequential scans. Segmenting lesions independently at each timepoint will lead to inconsistencies in lesion boundaries and in lesion detection, while modeling temporal dependencies via a change detection paradigm will provide a more consistent segmentation over time. We define new lesion volume between co-registered timepoints $t1$ and $t2$ as the volume of voxels that were not labelled as lesion at time $t1$ but were labelled as lesion at time $t2$. Similarily, we define resolving lesion volume as the volume of voxels that were labelled as lesion at time $t1$ but not at time $t2$. Table 1 shows new and resolving lesion volumes at w24, w36 and w48 using our method and using semi-manual lesion segmentations.

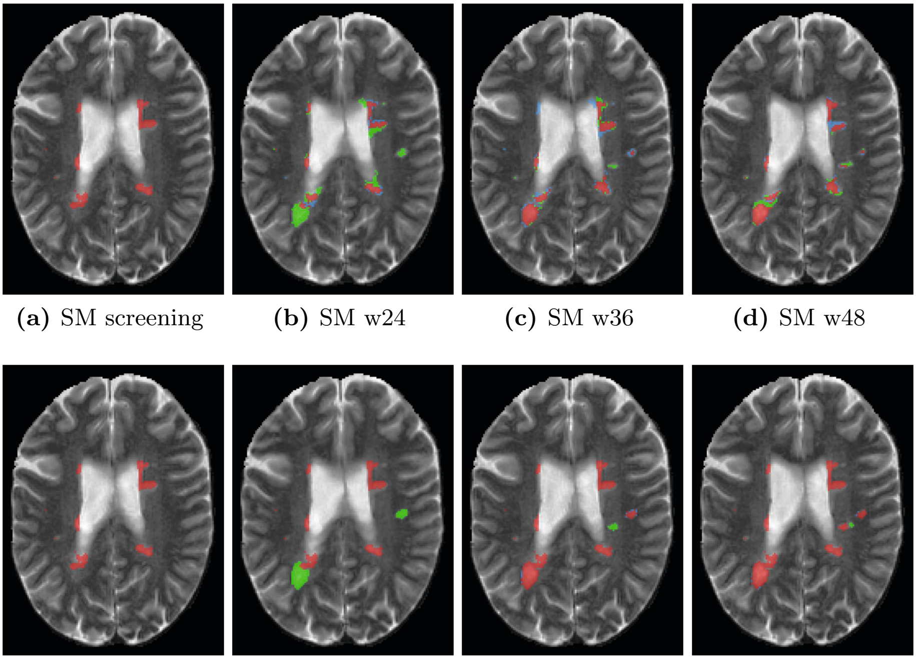
The proposed pipeline provides a much more temporally consistent segmentation of MS lesions, with the mean number of new and resolving lesion voxels detected over sequential timepoints both reduced by more than 90% as compared to semi-manual segmentations. Fig. 5 shows an example of a temporally consistent segmentation as generated by our method.

Effect Size Based on Measurements of Lesion Volume Change

Improved consistency of MS lesion segmentation is only useful if we can still detect veritable change. We demonstrate our sensitivity to change in lesion volume by calculating the effect size based on measurements of lesion volume change from screening to w48, as determined by our method and as determined from

Table 1. Lesion segmentation consistency as determined by mean volume of a) new lesion voxels (NV), and b) resolving lesion voxels (RV). Values represent mean volume in mm^3 ± 1 standard deviation, evaluated over all 73 subjects. Values are shown for change over subsequent timepoints (s-w24, w24-w36, w36-w48). SM = semi-manual lesion segmentations, CD = Proposed method based on change detection.

	s-w24		w24-w36		w36-w48	
	NV	RV	NV	RV	NV	RV
CD	301±1036	243±892	102±505	68±293	55±208	55±288
SM	2364±2223	2427±2534	2158±2104	2171±2053	2159±2017	2203±2092



(a) SM screening **(b)** SM w24 **(c)** SM w36 **(d)** SM w48

(e) screening **(f)** w24 **(g)** w36 **(h)** w48

Fig. 5. Example lesion segmentations over 4 timepoints. (a)-(d) shows the semi-manual (SM) reference and (e)-(h) shows our proposed pipeline based on change detection. The SM segmentation is used as a baseline segmentation for both methods. Stable portions of lesion are shown in red, new lesion voxels are shown in green and resolving in blue, all with respect to the previous timepoint. The proposed method shows increased segmentation consistency across time while remaining sensitive to real change.

the semi-manual lesion segmentations. For our method, volume change is determined for each subject by taking the difference between cumulative new lesion volume and cumulative resolving lesion volume over the four timepoints.

For reference semi-manual segmentations, volume change is determined by taking the volume derived from the reference segmentation at w48 and subtracting from the volume derived from the reference segmentation at screening.

The effect size is estimated by Cohen's *d* with a pooled standard deviation [13], and represents a normalized measure of difference between the treated and untreated groups. While both methods show a positive treatment effect (i.e. treated subjects are shown to have a smaller change in lesion volume than untreated), the calculated effect size is larger (ES=0.77) when based on lesion volume change measurements generated by our proposed method, as compared to semi-manual segmentations (ES=0.44). This suggests that our method remains sensitive to lesion change and provides greater statistical power to differentiate treated and untreated subjects, as shown graphically in Fig. 6.

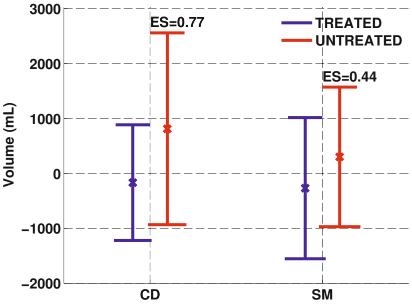


Fig. 6. Mean and standard deviation of lesion volume change from screening to w48 for treated (N=54) and untreated (N=19) subjects as measured from semi-manual lesion segmentations (SM) and the proposed method based on change detection (CD), along with corresponding effect size (ES) using Cohen's *d*.

4 Discussion

We have presented a novel method for detection of resolving MS lesion voxels in sequential brain MRI. By coupling our method with an existing method for detection of new MS lesions, we can provide a fully automated pipeline for determination of MS lesion volume change over serial scans based on change detection. Results demonstrate greater lesion segmentation consistency and improved statistical power to discriminate treatment arms using real clinical trial data, as compared to existing semi-manual segmentations. In addition, the ability to automatically detect resolving portions of MS lesions provides a potential measure of tissue repair, and as an aid for the analysis of MS lesion dynamics.

References

1. Meier, D.S., Weiner, H.L., Guttmann, C.R.: Time-series modeling of multiple sclerosis disease activity: a promising window on disease progression and repair potential? Neurotherapeutics 4(3), 485–498 (2007)

2. Meier, D.S., Guttmann, C.R.: MRI time series modeling of ms lesion development. Neuroimage 32(2), 531–537 (2006)
3. García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D., Louis Collins, D., Louis Collins, D.: Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. In: Medical Image Analysis (2012)
4. Lladó, X., Ganiler, O., Oliver, A., Martí, R., Freixenet, J., Valls, L., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À.: Automated detection of multiple sclerosis lesions in serial brain MRI. Neuroradiology 54(8), 787–807 (2012)
5. Elliott, C., Francis, S.J., Arnold, D.L., Collins, D.L., Arbel, T.: Bayesian classification of multiple sclerosis lesions in longitudinal MRI using subtraction images. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part II. LNCS, vol. 6362, pp. 290–297. Springer, Heidelberg (2010)
6. Elliott, C., et al.: Temporally consistent probabilistic detection of new multiple sclerosis lesions in brain MRI. IEEE TMI 32(8), 1490–1503 (2013)
7. Meier, D., Weiner, H., Guttmann, C.: MR imaging intensity modeling of damage and repair in multiple sclerosis: relationship of short-term lesion recovery to progression and disability. American Journal of Neuroradiology 28(10), 1956–1963 (2007)
8. Sled, J., Zijdenbos, A., Evans, A.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Transactions on Medical Imaging 17(1), 87–97 (1998)
9. Smith, S.: Fast robust automated brain extraction. Human Brain Mapping 17(3), 143–155 (2002)
10. Collins, D.L., Neelin, P., Peters, T.M., Evans, A.C.: Automatic 3D intersubject registration of MR volumetric data in standardized talairach space. Journal of computer assisted tomography 18(2), 192–205 (1994)
11. Nyul, L., Udupa, J., Zhang, X.: New variants of a method of MRI scale standardization. IEEE Transactions on Medical Imaging 19(2), 143–150 (2000)
12. Francis, S.: Automatic lesion identification in MRI of multiple sclerosis patients. Master's thesis, McGill University (2004)
13. Cumming, G.: Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. Routledge, New York (2012)

Author Index