SIAM J. Imaging Sciences
Vol. 10, No. 2, pp. 578–601

Most Likely Separation of Intensity and Warping Effects in Image Registration*

Line Kühnel†, Stefan Sommer†, Akshay Pai†, and Lars Lau Raket‡

Abstract. This paper introduces a class of mixed-effects models for joint modeling of spatially correlated intensity variation and warping variation in two-dimensional (2D) images. Spatially correlated intensity variation and warp variation are modeled as random effects, resulting in a nonlinear mixed-effects model that enables simultaneous estimation of template and model parameters by optimization of the likelihood function. We propose an algorithm for fitting the model which alternates estimation of variance parameters and image registration. This approach avoids the potential estimation bias in the template estimate that arises when treating registration as a preprocessing step. We apply the model to datasets of facial images and 2D brain magnetic resonance images to illustrate the simultaneous estimation and prediction of intensity and warp effects.

Key words. template estimation, image registration, separation of phase and intensity variation, nonlinear mixed-effects model

AMS subject classifications. 62-07, 62F99

DOI. 10.1137/16M1070980

1. Introduction.

When analyzing collections of imaging data, a general goal is to quantify similarities and differences across images. In medical image analysis and computational anatomy, a common goal is to find patterns that can distinguish morphologies of healthy and diseased subjects aiding the understanding of the population epidemiology. Such distinguishing patterns are typically investigated by comparing single observations to a representative member of the underlying population, and statistical analyses are performed relative to this representation. In the context of medical imaging, it has been customary to choose the template from the observed data as a common image of the population. However, such an approach has been shown to be highly dependent on the choice of the image. In more recent approaches, the templates are estimated using statistical methods that make use of the additional information provided by the observed data [19].

In order to quantify the differences between images, the dominant modes of variation in the data must be identified. Two major types of variability in a collection of comparable images are *intensity variation* and variation in *point correspondences*. Point correspondence or *warp* variation can be viewed as shape variability of an individual observation with respect to the template. Intensity variation is the variation that is left when the observations are

*Received by the editors April 19, 2016; accepted for publication (in revised form) March 13, 2017; published electronically April 27, 2017.
http://www.siam.org/journals/siims/10-2/M107098.html
Funding: The work of the authors was supported by the CSGB Centre for Stochastic Geometry and Advanced Bioimaging funded by a grant from the Villum foundation.
†Department of Computer Science, University of Copenhagen, Denmark (kuhnel@di.ku.dk, sommer@di.ku.dk, akshay@di.ku.dk).
‡Department of Mathematical Sciences, University of Copenhagen, Denmark (larslau@math.ku.dk).

Figure 1. *Fixed and random effects: The template (θ: leftmost) pertubed by random warp ($\theta \circ v$: second from left) and warp+spatially correlated intensity ($\theta \circ v + x$: third from left) together with independent noise ϵ constitute the observation (y: fourth from left). Right: The warp field v that brings the observation into spatial correspondence with θ overlayed the template. Estimation of template and model hyperparameters are conducted simultaneously with prediction of random effects allowing separation of the different factors in the nonlinear model.*

compensated for the true warp variation. This typically includes noise artifacts like systematic error and sensor noise or anatomical variation such as tissue density or tissue texture. Typically one would assume that the intensity variation consists of both independent noise and spatially correlated effects.

In this work, we introduce a flexible class of mixed-effects models that explicitly model the template as a fixed effect and intensity and warping variations as random effects; see Figure 1. This simultaneous approach enables separation of the random variation effects in a data-driven fashion using alternating maximum-likelihood estimation and prediction. The resulting model will therefore choose the separation of intensity and warping effects that is most likely given the patterns of variation found in the data. From the model specification and estimates, we are able to denoise observations through linear prediction in the model under the maximum likelihood estimates. Estimation in the model is performed with successive linearizations around the warp parameters enabling the use of linear mixed-effects predictors and avoiding the use of sampling techniques to account for nonlinear terms. We apply our method on datasets of face images and two-dimensional (2D) brain MRIs to illustrate its ability to estimate templates for populations and predict warp and intensity effects.

1.1. Outline of the paper. The paper is structured as follows. In section 2, we give an overview of previously introduced methods for analyzing image data with warp variation. Section 3 covers the mixed-effects model including a description of the estimation procedure (section 3.1) and how to predict from the model (section 3.2). In section 4, we give an example of how to model spatially correlated variations with a tied-down Brownian sheet. We consider two applications of the mixed-effects model to real-life datasets in section 5 and section 6 contains a simulation study that is used for comparing the precision of the model to more conventional approaches.

2. Background. The model introduced in this paper focuses on separately modeling the intensity and warp variation. Image registration conventionally only focuses on identifying

warp differences between pairs of images. The intensity variation is not included in the model and possible removal of this effect is considered as a pre- or post-processing step. The warp differences are often found by solving a variational problem of the form

$$E_{I_1,I_2}(\varphi) = R(\varphi) + \lambda S\left(I_1, I_2 \circ \varphi^{-1}\right); \tag{1}$$

see, for example, [39]. Here S measures the dissimilarity between the fixed image I_1 and the warped image $I_2 \circ \varphi^{-1}$, R is a regularization on the warp φ, and $\lambda > 0$ is a weight that is often chosen by ad hoc methods. After registration, either the warp, captured in φ, or the intensity differences between I_1 and $I_2 \circ \varphi^{-1}$ can be analyzed [40]. Several works have defined methods that incorporate registration as part of the defined models. The approach described in this paper will also regard registration as a part of the proposed model and adress the following three problems that arise in image analysis: (a) being able to estimate model parameters such as λ in a data-driven fashion; (b) assuming a generative statistical model that gives explicit interpretation of the terms that corresponds to the dissimilarity S and penalization R; and (c) simultaneously estimating population-wide effects such as the mean or template image and individual per-image effects, such as the warp and intensity effects. These features are of fundamental importance in image registration and many works have addressed combinations of them. The main difference of our approach to state-of-the-art statistical registration frameworks is that we propose a simultaneous random model for warp and intensity variation. As we will see, the combination of maximum likelihood estimation and the simultaneous random model for warp and intensity variation manifests itself in a trade-off where the uncertainty of both effects are taken into account simultaneously. As a result, when estimating fixed effects and predicting random effects in the model the most likely separation of the effects given the observed patterns of variation in the entire data material is used.

Methods for analyzing collections of image data, for example, template estimation in medical imaging [16], with both intensity and warping effects can be divided into two categories, *two-step methods* and *simultaneous methods*. Two-step methods perform alignment as a preprocessing step before analyzing the aligned data. Such methods can be problematic because the data are modified and the uncertainty related to the modification is ignored in the subsequent analysis. This means that the effect of intensity variation is generally underestimated, which can introduce bias in the analysis; see [34] for the corresponding situation in one-dimensional (1D) functional data analysis. Simultaneous methods, on the other hand, seek to analyze the images in a single step that includes the alignment procedure.

Conventional simultaneous methods typically use L^2 data terms to measure dissimilarity. Such dissimilarity measures are equivalent to the model assumption that the intensity variation in the image data consists solely of uncorrelated Gaussian noise. This approach is commonly used in image registration with the sum of squared differences dissimilarity measure, and in atlas estimation [48]. Since the L^2 data term is very fragile for systematic deviations from the model assumption, for example, contrast differences, the method can perform poorly. One solution to make the L^2 data term more robust against systematic intensity variation and, in general, to insufficient information in the data term is to add a strong penalty on the variation of the warping functions. This approach is, however, an implicit solution to the problem, since the gained robustness is a side effect of regularizing another model component.

As a consequence, the effect on the estimates is very hard to quantify, and it is very hard to specify a suitable regularization for a specific type of intensity variation. This approach is, for example, taken in the variational formulations of the template estimation problem in [16]. An elegant instance of this strategy is the Bayesian model presented in [1] where the warping functions are modeled as latent Gaussian effects with an unknown covariance that is estimated in a data-driven fashion. Conversely, systematic intensity variation can be sought to be removed prior to the analysis, in a reversed two-step method, for example, by using bias-correction techniques for MRI data [43]. The presence of warp variation can, however, influence the estimation of the intensity effects.

Analysis of images with systematic intensity differences can be improved using data dissimilarity measures that are robust or invariant to such systematic differences. However, robustness and invariance come at a cost in accuracy. By choosing a specific kind of invariance in the dissimilarity measure, the model is given a prespecified recipe for separating intensity and warping effects; the warps should maximize the invariant part of the residual under the given model parameters. Examples of classical robust data terms include L^1-norm data terms [31], Charbonnier data terms [4], and Lorentzian data terms [2]. Robust data terms are often challenging to use, since they may not be differentiable (L^1-norms) or may not be convex (Lorentzian data term). A wide variety of invariant data terms have been proposed, and are useful when the invariances represent a dominant mode of variation in the data. Examples of classical data terms that are invariant to various linear and nonlinear photometric relationships are normalized cross correlation, correlation ratio, and mutual information [20, 13, 36, 27]. Another approach for achieving robust or invariant data terms is to transform the data that are used in the data term. A classical idea is to match discretely computed gradients or other discretized derivative quantities [28]. A related idea is to construct invariant data terms based on discrete transformations. This type of approach has become increasingly popular in image matching in recent years. Examples include the rank transform and the census transform [47, 22, 10, 11], and more recently the complete rank transform [7]. While both robust and invariant data terms have been shown to give very good results in a wide array of applications, they induce a fixed measure of variation that does not directly model variation in the data. Thus, the general applicability of the method can come at the price of limited accuracy.

Several alternative approaches for analyzing warp and intensity simultaneously have been proposed [24, 15, 3, 45]. In [24] warps between images are considered as a combination of two transformation fields, one representing the image motion (warp effect) and one describing the change of image brightness (intensity effect). Based on this definition warp and intensity variations can be modeled simultaneously. An alternative approach is considered in [15], where an invariant metric is used, which enables analysis of the dissimilarity in point correspondences between images disregarding the intensity variation. These methods are not statistical in the sense that they do not seek to model the random structures of the variation of the image data. A statistical model is presented in [3], where parameters for texture, shape variation (warp), and rendering are estimated using maximizing-a-posteriori estimation.

To overcome the mentioned limitations of conventional approaches, we propose to do statistical modeling of the sources of variation in data. By using a statistical model where we assume parametric covariance structures for the different types of observed variation, the

variance parameters can be estimated from the data. The contribution of different types of variation is thus weighted differently in the data term. By using, for example, maximum-likelihood estimation, the most likely form of the variation given the data is penalized the least. We emphasize that in contrast to previous mixed-effects models incorporating warp effects [1, 48], the goal here is to simultaneously model warp and intensity effects. These effects impose randomness relative to a template, the fixed-effect, that is estimated during the inference process.

The nonlinear mixed-effects models are a commonly used tool in statistics. These types of models can be computationally intensive to fit, and are rarely used for analyzing large data sizes such as image data. We formulate the proposed model as a nonlinear mixed-effects model and demonstrate how certain model choices can be used to make estimation in the model computationally feasible for large data sizes. The model incorporates random intensity and warping effects in a small-deformation setting: We do not require warping functions to produce diffeomorphisms. The geometric structure is therefore more straightforward than in, for example, the large deformation diffeomorphic metric mapping (LDDMM) model [46]. From a statistical perspective, the small-deformation setting is much easier to handle than the large-deformation setting where warping functions are restricted to produce diffeomorphisms.

Instead of requiring diffeomorphisms, we propose a class of models that will produce warping functions that in most cases do not fold. Another advantage of the small-deformation setting is that we can model the warping effects as latent Gaussian disparity vectors in the domain. Such direct modeling allows one to compute a high-quality approximation of the likelihood function by linearizing the model around the modes of the nonlinear latent random variables. The linearized model can be handled using conventional methods for linear mixed-effects models [29] which are very efficient compared to sampling-based estimation procedures.

In the large-deformation setting, the metamorphosis model [41, 42] extends the LDDMM framework for image registration [46] to include intensity change in images. Warp and intensity differences are modeled separately in metamorphosis with a Riemannian structure measuring infinitesimal variation in both warp and intensity. While this separation has similarities to the statistical model presented here, we are not aware of any works which have considered likelihood-based estimation of variables in metamorphosis models.

3. Statistical model.

We consider spatial functional data defined on $\mathbb{R}^2$ taking values in $\mathbb{R}$. Let $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ be n functional observations on a regular lattice with $m = m_1 m_2$ points (s_j, t_k), that is, $\boldsymbol{y}_i = (y_i(s_j, t_k))_{j,k}$ for $j = 1, \dots, m_1$, $k = 1, \dots, m_2$. Consider the model in the image space

$$y_i(s_j, t_k) = \theta(v_i(s_j, t_k)) + x_i(s_j, t_k) + \varepsilon_{ijk}, \tag{2}$$

for $i = 1, \dots, n$, $j = 1, \dots, m_1$, and $k = 1, \dots, m_2$. Here $\theta \colon \mathbb{R}^2 \to \mathbb{R}$ denotes the template and $v_i \colon \mathbb{R}^2 \to \mathbb{R}^2$ is a warping function matching a point in y to a point in the template θ. Moreover x_i is the random spatially correlated intensity variation for which we assume that $\boldsymbol{x}_i = (x_i(s_j, t_k))_{j,k} \sim \mathcal{N}(0, \sigma^2 S)$, where the spatial correlation is determined by the covariance matrix S. The term $\varepsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ models independent noise. The template θ is a fixed effect while v_i, x_i, and ε_{ijk} are random.

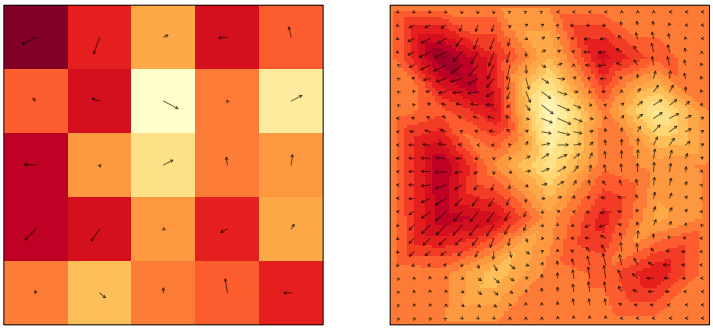

Figure 2. *An example of disparity vectors at a* 5×5 *grid of anchor points and the corresponding warping function.*

We will consider warping functions of the form

$$v_i(s,t) = v(s,t,\boldsymbol{w}_i) = \begin{pmatrix} s \\ t \end{pmatrix} + \mathcal{E}_{\boldsymbol{w}_i}(s,t),$$

where $\mathcal{E}_{\boldsymbol{w}_i} \colon \mathbb{R}^2 \to \mathbb{R}^2$ is a coordinatewise bilinear spline interpolation of $\boldsymbol{w}_i \in \mathbb{R}^{m_w^1 \times m_w^2 \times 2}$ on a lattice spanned by $\boldsymbol{s}_w \in \mathbb{R}^{m_w^1}, \boldsymbol{t}_w \in \mathbb{R}^{m_w^2}$. In other words, $\boldsymbol{w}_i$ models discrete spatial displacements at the lattice anchor points. Figure 2 shows an example of disparity vectors on a grid of anchor points and the corresponding warping function.

The displacements are modeled as random effects, $\boldsymbol{w}_i \sim \mathcal{N}(0, \sigma^2 C)$, where C is a $2m_w^1 m_w^2 \times 2m_w^1 m_w^2$ covariance matrix and, as a result, the warping functions can be considered nonlinear functional random effects. As $\boldsymbol{w}_i$ is assumed to be normally distributed with mean zero, small displacements are favored and, hence, the warp effect will be less prone to fold. The model is a spatial extension of the phase and amplitude varying population pattern (PAVPOP) model for curves [34, 32].

3.1. Estimation. First, we will consider estimation of the template θ from the functional observations, and we will estimate the contributions of the different sources of variation. In the proposed model, this is equivalent to estimating the covariance structure C for the warping parameters, the covariance structure S for the spatially correlated intensity variation, and the noise variance σ^2. The estimate of the template is found by considering model (2) in the back-warped template space

$$y_i\left(v_i^{-1}(s_j,t_k)\right) = \theta(s_j,t_k) + x_i\left(v_i^{-1}(s_j,t_k)\right) + \tilde{\varepsilon}_{ijk}. \tag{3}$$

Because every back-warped image represents θ on the observation lattice, a computationally attractive parametrization is to model θ using one parameter per observation point, and evaluate nonobservation points using bilinear interpolation. This parametrization is attractive, because Henderson's mixed-model equations [12, 35] suggest that the conditional estimate for $\theta(s_j,t_k)$ given $\boldsymbol{w}_1,\ldots,\boldsymbol{w}_n$ is the pointwise average

$$\hat{\theta}(s_j,t_k) = \frac{1}{n}\sum_{i=1}^{n} y_i\left(v_i^{-1}(s_j,t_k)\right), \tag{4}$$

if we ignore the slight change in covariance resulting from the back warping of the random intensity effects. As this estimator depends on the warping parameters, the estimation of θ and the variance parameters have to be performed simultaneously with the prediction of the warping parameters. We note that, as in any linear model, the estimate of the template is generally quite robust against slight misspecifications of the covariance structure. Also, the idea of estimating the template conditional on the posterior warp is similar to the idea of using a hard EM algorithm for computing the maximum-likelihood estimator for θ [23].

We use maximum-likelihood estimation to estimate variance parameters, that is, we need to minimize the negative log-likelihood function of model (2). Note that (2) contains nonlinear random effects due to the term $\theta(v_i(s,t,\boldsymbol{w}_i))$, where $\theta \circ v_i$ is a nonlinear transformation of $\boldsymbol{w}_i$. We handle the nonlinearity and approximate the likelihood by linearizing the model (2) around the current predictions $\boldsymbol{w}_i^0$ of the warping parameters $\boldsymbol{w}_i$:

$$\begin{aligned} y_i(s_j,t_k) &\approx \theta\left(v\left(s_j,t_k,\boldsymbol{w}_i^0\right)\right) \\ &\quad + \left(\nabla\theta\left(v\left(s_j,t_k,\boldsymbol{w}_i^0\right)\right)\right)^\top J_{\boldsymbol{w}_i} v(s_j,t_k,\boldsymbol{w}_i)\Big|_{\boldsymbol{w}_i=\boldsymbol{w}_i^0}\left(\boldsymbol{w}_i-\boldsymbol{w}_i^0\right) \\ &\quad + x_i(s_j,t_k)+\varepsilon_{ijk} \\ &= \theta\left(v\left(s_j,t_k,\boldsymbol{w}_i^0\right)\right) + Z_{ijk}\left(\boldsymbol{w}_i-\boldsymbol{w}_i^0\right) + x_i(s_j,t_k)+\varepsilon_{ijk}, \end{aligned} \tag{5}$$

where $J_{\boldsymbol{w}_i} v(s_j,t_k,\boldsymbol{w}_i)$ denotes the Jacobian matrix of v with respect to $\boldsymbol{w}_i$ and

$$Z_{ijk} = \left(\nabla\theta\left(v\left(s_j,t_k,\boldsymbol{w}_i^0\right)\right)\right)^\top J_{\boldsymbol{w}_i} v(s_j,t_k,\boldsymbol{w}_i)\Big|_{\boldsymbol{w}_i=\boldsymbol{w}_i^0}. \tag{6}$$

Letting $Z_i = (Z_{ijk})_{jk} \in \mathbb{R}^{m\times 2m_w^1 m_w^2}$, the linearized model can be rewritten

$$\boldsymbol{y}_i \approx \boldsymbol{\theta}^{\boldsymbol{w}_i^0} + Z_i\left(\boldsymbol{w}_i-\boldsymbol{w}_i^0\right) + \boldsymbol{x}_i + \boldsymbol{\varepsilon_i}. \tag{7}$$

We notice that, in this manner, $\boldsymbol{y}_i$ can be approximated as a linear combination of normally distributed variables, hence, the negative log-likelihood for the linearized model is given by

$$\begin{aligned} \ell_{\boldsymbol{y}}\left(\theta,C,\sigma^2\right) &= \frac{nm_1m_2}{2}\log\sigma^2 + \frac{1}{2}\sum_{i=1}^n \log\det V_i \\ &\quad + \frac{1}{2\sigma^2}\sum_{i=1}^n \left(\boldsymbol{y}_i - \boldsymbol{\theta}^{\boldsymbol{w}_i^0} + Z_i\boldsymbol{w}_i^0\right)^\top V_i^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\theta}^{\boldsymbol{w}_i^0} + Z_i\boldsymbol{w}_i^0\right), \end{aligned} \tag{8}$$

where $V_i = Z_i C Z_i^\top + S + \mathbb{I}_m$. The idea of linearizing nonlinear mixed-effects models in the nonlinear random effects is a solution that has been shown to be effective and which is implemented in standard software packages [18, 29, 30]. The proposed model is, however, both more general and computationally demanding than what can be handled by conventional software packages. Furthermore, we note that the linearization in a random effect as done in model (7) is fundamentally different than the conventional linearization of a nonlinear dissimilarity measure such as in the variational problem (1). As we see from the linearized model (7), the density of $\theta(v(s_j,t_k,\boldsymbol{w}_i)$ is approximated by the density of a linear combination, $\theta(v(s_j,t_k,\boldsymbol{w}_i^0)) + Z_{ijk}(\boldsymbol{w}_i-\boldsymbol{w}_i^0)$, of multivariate Gaussian variables. The likelihood function for the first-order Taylor expansion in $\boldsymbol{w}_i$ of the model (2) is thus a Laplace approximation of the true likelihood, and the quality of this approximation is approximately second order [44].

3.1.1. Computing the likelihood function. As mentioned above the proposed model is computationally demanding. Even the approximated likelihood function given in (8) is not directly computable because of the large data sizes. In particular, the computations related to determinants and inverses of the covariance matrix V_i are infeasible unless we impose certain structures on these. In the following, we will assume that the covariance matrix for the spatially correlated intensity variation S has full rank and sparse inverse. We stress that this assumption is merely made for computational convenience and that the proposed methodology is also valid for nonsparse precision matrices. The zeros in the precision matrix S^{-1} are equivalent to assuming conditional independences between the intensity variation in corresponding pixels given all other pixels [17]. A variety of classical models have this structure, in particular (higher-order) Gaussian Markov random fields models have sparse precision matrices because of their Markov property.

To efficiently do computations with the covariances $V_i = Z_i C Z_i^\top + S + \mathbb{I}_m$, we exploit the structure of the matrix. The first term $Z_i C Z_i^\top$ is an update to the intensity covariance $S + \mathbb{I}_m$ with a maximal rank of $2m_w^1 m_w^2$. Furthermore, the first term of the intensity covariance S has a sparse inverse and the second term $\mathbb{I}_m$ is, of course, sparse with a sparse inverse. Using the Woodbury matrix identity, we obtain

$$\begin{aligned} V_i^{-1} &= \left(Z_i C Z_i^\top + S + \mathbb{I}_m\right)^{-1} \\ &= (S + \mathbb{I}_m)^{-1} - (S + \mathbb{I}_m)^{-1} Z_i \left(C^{-1} + Z_i^\top (S + \mathbb{I}_m)^{-1} Z_i\right)^{-1} Z_i^\top (S + \mathbb{I}_m)^{-1} \end{aligned}$$

which can be computed if we can efficiently compute the inverse of the potentially huge $m \times m$ intensity covariance matrix $(S + \mathbb{I}_m)^{-1}$. We can rewrite the inverse intensity covariance as

$$(S + \mathbb{I}_m)^{-1} = \mathbb{I}_m - \left(\mathbb{I}_m + S^{-1}\right)^{-1}.$$

Thus we can write V_i^{-1} in a way that only involves operations on sparse matrices. To compute the inner product $\boldsymbol{y}^\top V_i^{-1} \boldsymbol{y}$, we first form the matrix $\mathbb{I}_m + S^{-1}$ and compute its Cholesky decomposition using the Ng–Peyton method [25] implemented in the `spam` R-package [9]. By solving a low-rank linear system using the Cholesky decomposition, we can thus compute $L = (C^{-1} + Z_i^\top (S + \mathbb{I}_m)^{-1} Z_i)^{-1}$. The inner product is then efficiently computed as

$$\boldsymbol{y}^\top V_i^{-1} \boldsymbol{y} = \boldsymbol{y}^\top \boldsymbol{x} - (Z_i \boldsymbol{x})^\top L Z_i \boldsymbol{x},$$

where

$$\boldsymbol{x} = (S + \mathbb{I}_m)^{-1} \boldsymbol{y}.$$

To compute the log determinant in the likelihood, one can use the matrix determinant lemma similarly to what was done above to split the computations into low-rank computations and computing the determinant of $S + \mathbb{I}_m$,

$$\det(V_i) = \det\left(Z_i C Z_i^\top + S + \mathbb{I}_m\right) = \det\left(C^{-1} + Z_i^\top (S + \mathbb{I}_m)^{-1} Z_i\right) \det(C) \det(S + \mathbb{I}_m).$$

For the models that we will consider, the latter computation is done by using the operator approximation proposed in [33] which, for image data with sufficiently high resolution (e.g., $m > 30$), gives a high-quality approximation of the determinant of the intensity covariance that can be computed in constant time.

By taking the described strategy, we never need to form a dense $m \times m$ matrix, and we can take advantage of the sparse and low-rank structures to reduce the computation time drastically. Furthermore, the fact that we assume equal-size images allows us to only do a single Cholesky factorization per likelihood computation, which is further accelerated by using the updating scheme described in [25].

3.2. Prediction. After the maximum-likelihood estimation of the template θ and the variance parameters, we have an estimate for the distribution of the warping parameters. We are therefore able to predict the warping functions that are most likely to have occurred given the observed data. This prediction parallels the conventional estimation of deformation functions in image registration. Let $p_{w_i|y_i}$ be the density for the distribution of the warping functions given the data and define p_{w_i}, $p_{y_i|w_i}$ in a similar manner. Then, by applying $p_{w_i|y_i} \propto p_{y_i|w_i} p_{w_i}$, we see that the warping functions that are most likely to occur are the minimizers of the posterior

$$-\log(p_{w_i|y_i}) \propto \frac{1}{2\sigma^2}(\boldsymbol{y}_i - \boldsymbol{\theta}^{\boldsymbol{w}_i})^\top (S + I_m)^{-1}(\boldsymbol{y}_i - \boldsymbol{\theta}^{\boldsymbol{w}_i}) + \frac{1}{2\sigma^2}\boldsymbol{w}_i^\top C^{-1}\boldsymbol{w}_i. \tag{9}$$

Given the updated predictions $\hat{\boldsymbol{w}}_i$ of the warping parameters, we update the estimate of the template and then minimize the likelihood (8) to obtain updated estimates of the variances. This procedure is then repeated until convergence is obtained. The estimation algorithm is given in Algorithm 1. The run times for the algorithm will be very different depending on the data in question. As an example, we ran the model for 10 MRI midsaggital slices (for more details, see section 5.2) of size 210×210, with $i_{\max} = 5, j_{\max} = 3$. We ran the algorithm on an Intel Xeon E5-2680 2.5 GHz processor. The run time needed for full maximum likelihood estimation in this setup was 1 hour and 15 minutes using a single core. This run time is without parallization, but it is possible to apply parallization to make the algorithm go faster.

The spatially correlated intensity variation can also be predicted. Either as the best linear unbiased prediction $\mathrm{E}[\boldsymbol{x}_i \,|\, \boldsymbol{y}]$ from the linearized model (7) (see, e.g., [21, (5)]). Alternatively, to avoid a linear correction step when predicting $\boldsymbol{w}_i$, one can compute the best linear unbiased prediction given the maximum-a-posteriori warp variables

$$\mathrm{E}[x_i(s,t) \,|\, \boldsymbol{y}_i, \boldsymbol{w}_i = \hat{\boldsymbol{w}}_i] = S(S + I_m)^{-1}\left(\boldsymbol{y}_i - \hat{\boldsymbol{\theta}}^{\hat{\boldsymbol{w}}_i}\right). \tag{10}$$

The prediction of the spatially correlated intensity variation can, for example, be used for bias field correction of the images.

4. Models for the spatially correlated variations. The main challenge of the presented methods is the computability of the likelihood function, in particular, computations related to the $m \times m$ covariance matrix of the spatially correlated intensity variation S. The same issues are not associated with the covariance matrix C, for the warping parameters, as the dimensions of this matrix are considerably smaller than the dimensions of S. At the end of this section, we will give a short description of how the displacement vectors can be modeled, but first we consider the covariance matrix S.

As mentioned in the previous section, the path we will pursue to make likelihood computations efficient is to assume that the systematic random effect $\boldsymbol{x}_i$ has a covariance matrix S with sparse inverse. In particular, modeling $\boldsymbol{x}_i$ as a Gaussian Markov random field will

Algorithm 1 Inference in the model (2).

Data: $\boldsymbol{y}$

Result: Estimates of the fixed effect and variance parameters of the model, and the resulting predictions of the warping parameters $\boldsymbol{w}$

```
// Initialize parameters
```
Initialize $\boldsymbol{w}^0$

Compute $\hat{\boldsymbol{\theta}}^{\boldsymbol{w}^0}$ following (4)

for $i = 1$ *to* $i_{\max}$ **do**

```
    // Outer loop:  parameters
```
    Estimate variance parameters by minimizing (8)

    for $j = 1$ *to* $j_{\max}$ **do**

```
        // Inner loop:  fixed effect, warping parameters
```
        Predict warping parameters by minimizing (9)

        Update linearization points $\boldsymbol{w}^0$ to current prediction

        Recompute $\hat{\boldsymbol{\theta}}^{\boldsymbol{w}^0}$ from (4)

    end

end

give sparse precision matrices S^{-1}. The Markov random field structure gives a versatile class of models that has been demonstrated to be able to approximate the properties of general Gaussian fields surprisingly well [37]. Estimation of a sparse precision matrix is a fundamental problem and a vast literature exists on the subject. We mention in passing the fundamental works, [5, 8], which could be adapted to the present setup to estimate unstructured sparse precision matrices. We will, however, not pursue that extension in the present paper.

We here model $\boldsymbol{x}_i$ as a tied-down Brownian sheet, which is the generalization of the Brownian bridge (which is Markov) to the unit square $[0,1]^2$. The covariance function, $\mathcal{S}\colon [0,1]^2 \times [0,1]^2 \to \mathbb{R}$, for the tied-down Brownian sheet is

$$\mathcal{S}((s,t),(s',t')) = \tau^2(s \wedge s' - ss')(t \wedge t' - tt'), \qquad \tau > 0.$$

The covariance is 0 along the boundary of the unit square and reaches its maximal variance at the center of the image. These properties seem reasonable for many image analysis tasks, where one would expect the subject matter to be centered in the image with little or no variation along the image boundary.

Let S be the covariance matrix for a Brownian sheet observed at the lattice spanned by $(s_1, \dots, s_{m_1})$ and $(t_1, \dots, t_{m_2})$, $s_i = i/(m_1+1)$, $t_i = i/(m_2+1)$, with row-major ordering. The precision matrix S^{-1} is sparse with the following structure for points corresponding to nonboundary elements:

$$\frac{1}{\tau^2(m_1+1)(m_2+1)} S^{-1}[i,j] = \begin{cases} 4 & \text{if } j = i, \\ -2 & \text{if } j \in \{i-1, i+1, i+m_2, i-m_2\}, \\ 1 & \text{if } j \in \{i-1-m_2, i+1-m_2, i-1+m_2, i+1+m_2\}. \end{cases}$$

For boundary elements, the j elements outside the observation boundary vanish.

As explained in section 3.1.1, the computational difficulties related to the computation of the log determinant in the negative log-likelihood function (8) comes down to computing the log determinant of the intensity covariance $S + \mathbb{I}_m$. For the tied-down Brownian sheet, the log determinant can be approximated by means of the operator approximation given in [33, Example 3.4]. The approximation is given by

$$\log\det(S+\mathbb{I}_m) = \sum_{\ell=1}^{\infty} \log\left(\frac{\pi\ell}{\sqrt{\tau^2(m_1+1)(m_2+1)}}\sinh\left(\frac{\sqrt{\tau^2(m_1+1)(m_2+1)}}{\pi\ell}\right)\right).$$

To compute the approximation we cut the sum off after 10,000 terms.

As a final remark, we note that the covariance function $\tau^{-2}\mathcal{S}$ is the Green's function for the differential operator $\partial_s^2\partial_t^2$ on $[0,1]^2$ under homogeneous Dirichlet boundary conditions. Thus the conditional linear prediction of $\boldsymbol{x}_i$ given by (10) is equivalent to estimating the systematic part of the residual as a generalized smoothing spline with roughness penalty

$$\frac{1}{2\tau^2}\int_0^1\int_0^1 x_i(s,t)\partial_s^2\partial_t^2 x_i(s,t)\,\mathrm{d}s\,\mathrm{d}t = \frac{1}{2\tau^2}\int_0^1\int_0^1 \|\partial_s\partial_t x_i(s,t)\|^2\,\mathrm{d}s\,\mathrm{d}t.$$

The tied-down Brownian sheet can also be used to model the covariance between the displacement vectors. Here the displacement vectors given by the warping variables $\boldsymbol{w}_i$ are modeled as discretely observed tied-down Brownian sheets in each displacement coordinate. As was the case for the intensity covariance, this model is a good match to image data since it allows the largest deformations around the middle of the image. Furthermore, the fact that the model is tied down along the boundary means that we will predict the warping functions to be the identity along the boundary of the domain $[0,1]^2$, and for the found variance parameters, the predicted warping functions will be homeomorphic maps of $[0,1]^2$ onto $[0,1]^2$ with high probability.

In the applications in the next section, we will use the tied-down Brownian sheet to model the spatially correlated variations.

5. Applications.

In this section, we will apply the developed methodology on two different real-life datasets. In the first example, we apply the model to a collection of face images that are difficult to compare due to varying expressions and lighting sources. We compare the results of the proposed model to conventional registration methods and demonstrate the effects of the simultaneous modeling of intensity and warp effects. In the second example, we apply the methodology to the problem of estimating a template from affinely aligned 2D MR images of brains.

5.1. Face registration.

Consider the ten 92×112 face images from the AT&T Laboratories Cambridge Face Database [38] in Figure 3. The images are all of the same person, but vary in head position, expression, and lighting. The dataset contains two challenges from a registration perspective, namely, the differences in expression that cause disocclusions or occlusions (e.g., showing teeth, closing eyes) resulting in large local deviations, and the difference in placement of the lighting source that causes strong systematic deviations throughout the face.

Figure 3. *Ten images of the same face with varying expressions and illumination. The images are from the AT&T Laboratories Cambridge Face Database* [38].

To estimate a template face from these images, the characteristic features of the face should be aligned, and the systematic local and global deviations should be accounted for. In the proposed model (2), these deviations are explicitly modeled through the random effect x_i.

Using the maximum-likelihood estimation procedure, we fitted the model to the data using displacement vectors $\boldsymbol{w}_i$ on an equidistant 4×4 interior grid in $[0,1]^2$. We used 5 outer and 3 inner iterations in Algorithm 1. The image value range was scaled to $[0,1]$. The estimated variance scale for the random effect x_i was $\hat{\sigma}^2\hat{\tau}^2 = 0.658$; for the warp variables, the variance scale was estimated to be $\hat{\sigma}^2\hat{\gamma}^2 = 0.0680$; and for the residual variance, the estimated scale was $\hat{\sigma}^2 = 0.00134$.

To illustrate the effect of the simultaneous modeling of random intensity and warp effects, we estimated a face template using three more conventional variants of the proposed framework: a pointwise estimation that corresponds to model (2) with no warping effect; a *Procrustes* model that corresponds to model (2) with no intensity component and where the warp variables $\boldsymbol{w}_i$ were modeled as unknown parameters and estimated using maximum-likelihood estimation; and a *warp-regularized Procrustes* method where the warp variables $\boldsymbol{w}_i$ were penalized using a term $\lambda \boldsymbol{w}_i^\top C^{-1} \boldsymbol{w}_i$, where C^{-1} is the precision matrix for the 2D tied-down Brownian sheet with smoothing parameter $\lambda = 3.125$ (chosen to give good visual results).

The estimated templates for the proposed model and the alternative models described above can be found in Figure 4. Going from left to right, it is clear that the sharpness and representativeness of the estimates increase.

To validate the models, we can consider how well they predict the observed faces under the maximum-likelihood estimates and posterior warp predictions. These predictions are displayed in Figure 5. The rightmost column displays the five most deviating observed faces. From the left, the first three columns show the corresponding predictions from the Procrustes model, the

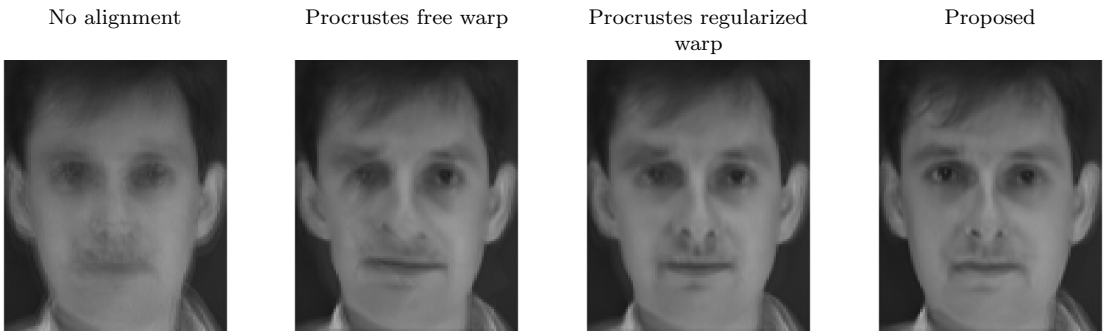

Figure 4. *Estimates for the fixed effect θ using different models. The models used to calculate the estimates are, from left to right, model assuming no warping effect and Gaussian white noise for the intensity model, the same model but with a free warping function based on 16 displacement vectors, the same model but with a penalized estimation of warping functions (2D tied-down Brownian sheet with scale fixed $\tau = 0.4$), the full model (2).*

warp-regularized Procrustes model, and, for comparison, the predicted warped templates from the proposed model. It is clear that both the sharpness and the representativeness increase from left to right. The predictions in the third column show the warped template of model (2) which does not include the predicted intensity effect x_i. The fourth column displays the full prediction from the proposed model given as the best linear unbiased prediction conditional on the maximum-a-posteriori warp variables $\hat{\theta}(v(s,t,\hat{\boldsymbol{w}}_i)) + \mathrm{E}[x_i(s,t) \mid \boldsymbol{y}_i, \boldsymbol{w}_i = \hat{\boldsymbol{w}}_i]$. The full predictions are very faithful to the observations, with only minor visible deviations around the eyes in the second and fifth row. This suggests that the chosen model for the spatially correlated intensity variation, the tied-down Brownian sheet, is sufficiently versatile to model the systematic part of the residuals.

5.2. MRI slices. The data considered in this section are based on three-dimensional (3D) MR images from the ADNI database [26]. We have based the example on 50 images with 18 normal controls, 13 with Alzheimer's disease, and 19 who are mild cognitively impaired. The 3D images were initially affinely aligned with 12 degrees of freedom and normalized mutual information as a similarity measure. After the registration, the midsagittal slices were chosen as observations. Moreover the images were intensity normalized to $[0,1]$ and afterwards the midsagittal plane was chosen as the final observations. The 50 midsagittal planes are given as 210×210 observations on an equidistant grid on $[0,1]^2$. Six samples are displayed in Figure 6 where differences in both contrast, placement, and shape of the brains are apparent.

For the given data, we used 25 displacement vectors $\boldsymbol{w}_i$ on an equidistant 5×5 interior grid in $[0,1]^2$. The number of inner iterations in the algorithm was set to 3, while the number of outer iterations was set to 5 as the variance parameters and likelihood value already stabilized after a couple of iterations. The estimated variance scales are given by $\hat{\sigma}^2\hat{\tau}^2 = 2.23$ for the spatially correlated intensity variation, $\hat{\sigma}^2\hat{\gamma}^2 = 0.202$ for the warp variation, and $\hat{\sigma}^2 = 7.79 \cdot 10^{-4}$ for the residual variance. The estimated template can be found in the rightmost column in Figure 7.

For comparison, we have estimated a template without any additional warping (i.e., only using the rigidly aligned slices), and a template estimated using a Procrustes model with

Figure 5. *Model predictions of five face images (rightmost column). The two first columns display the maximum-likelihood predictions from the Procrustes and regularized Procrustes models. The third column displays the warped template $\hat{\theta}(v(s,t,\hat{\boldsymbol{w}}_i))$, where $\hat{\boldsymbol{w}}_i$ is the most likely warp given data $\boldsymbol{y}$. The fourth column displays the full conditional prediction given the posterior warp variables $\hat{\theta}(v(s,t,\hat{\boldsymbol{w}}_i))+\mathrm{E}[x_i(s,t)\,|\,\boldsymbol{y}_i,\boldsymbol{w}_i=\hat{\boldsymbol{w}}_i]$.*

fixed warping effects and no systematic intensity variation, but otherwise comparable to the proposed model. These templates can be found in the leftmost and middle columns of Figure 7. Comparing the three, we see a clear increase in details and sharpness from left to right. The reason for the superiority of the proposed method is both that the regularization of warps is based on maximum-likelihood estimation of variance parameters, but also that the prediction of warps takes the systematic deviations into account. Indeed, we can rewrite the data term

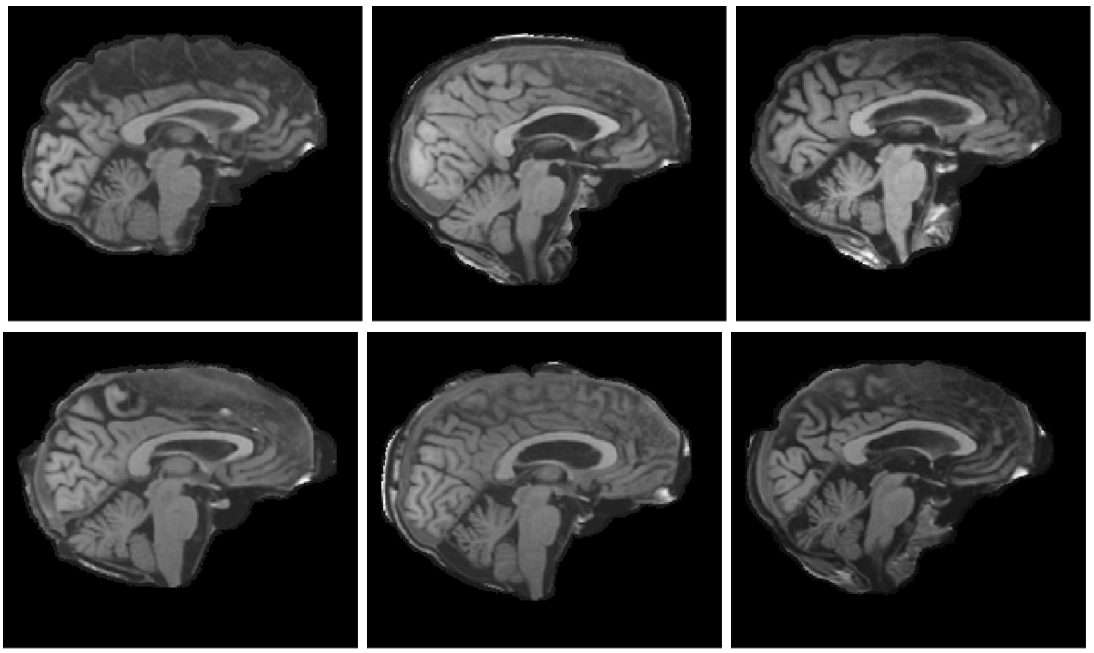

Figure 6. *A sample of six MRI slices from the data set of* 50 *midsagittal MRI slices.*

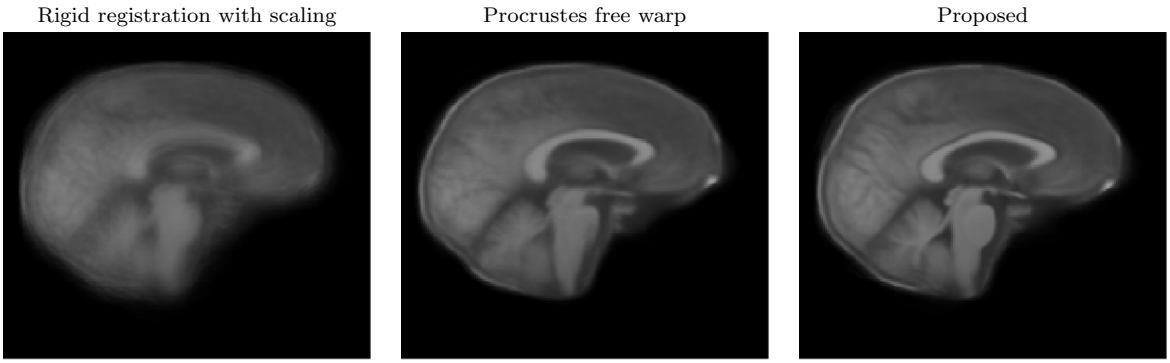

Figure 7. *Estimates for the fixed effect θ in three different models. From left to right: pointwise mean after rigid registration and scaling; nonregularized Procrustes; and the proposed model* (2).

in the posterior (9) as

$$(\boldsymbol{y}_i - \boldsymbol{\theta}^{\boldsymbol{w}_i} - \mathrm{E}[\boldsymbol{x}_i \,|\, \boldsymbol{y}_i, \boldsymbol{w}_i])^\top (\boldsymbol{y}_i - \boldsymbol{\theta}^{\boldsymbol{w}_i} - \mathrm{E}[\boldsymbol{x}_i \,|\, \boldsymbol{y}_i, \boldsymbol{w}_i]) + \mathrm{E}[\boldsymbol{x}_i \,|\, \boldsymbol{y}_i, \boldsymbol{w}_i]^\top S^{-1} \mathrm{E}[\boldsymbol{x}_i \,|\, \boldsymbol{y}_i, \boldsymbol{w}_i].$$

Thus, in the prediction of warps, there is a trade-off between the regularity of the displacement vectors (the term $\boldsymbol{w}_i^\top C^{-1} \boldsymbol{w}_i$ in (9)) and the regularity of the predicted spatially correlated intensity variation given the displacement vectors (the term $\mathrm{E}[\boldsymbol{x}_i \,|\, \boldsymbol{y}_i, \boldsymbol{w}_i]^\top S^{-1} \mathrm{E}[\boldsymbol{x}_i \,|\, \boldsymbol{y}_i, \boldsymbol{w}_i]$).

The difference in regularization of the warps is shown in Figure 8, where the estimated warps using the Procrustes model are compared to the predicted warps from the proposed

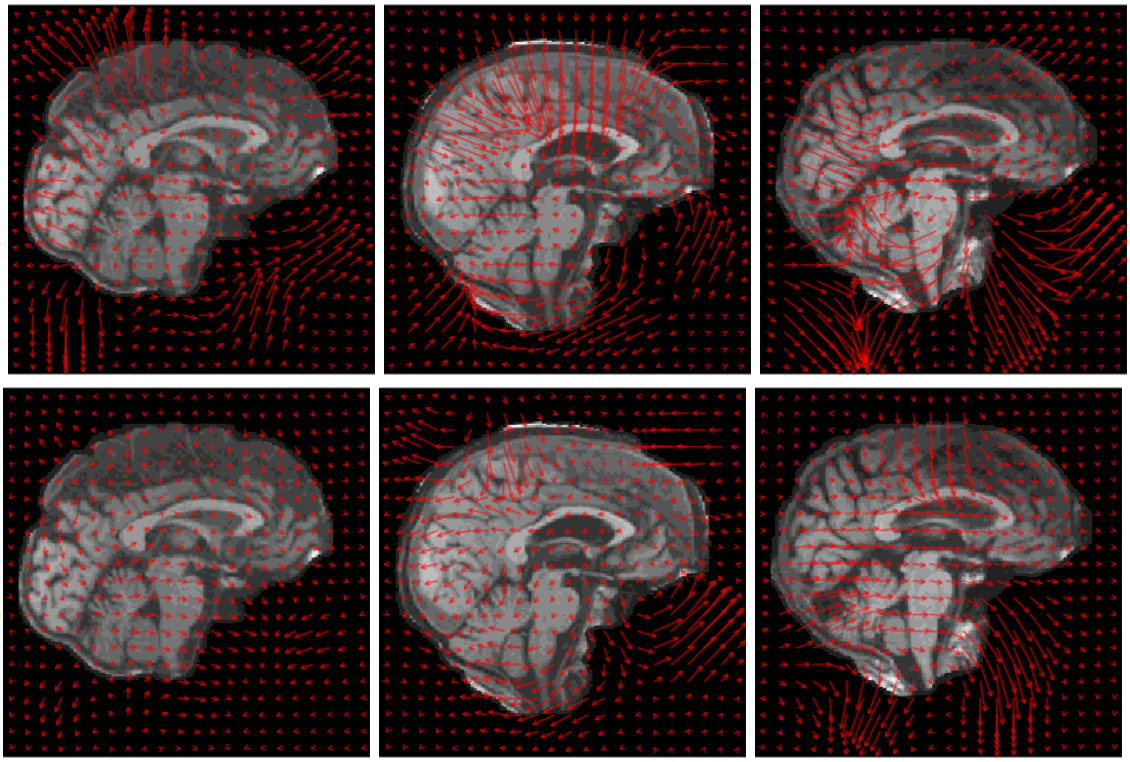

Figure 8. *Three MRI slices and their estimated/predicted warping functions for the Procrustes model and the proposed model. The top row shows the Procrustes displacement fields, while the displacement fields for the proposed model are given in the bottom row. The arrows correspond to the deformation of the observation to the template.*

model. We see that the proposed model predicts much smaller warps than the Procrustes model.

One of the advantages of the mixed-effects model is that we are able to predict the systematic part of the intensity variation of each image, which in turn also gives a prediction of the residual intensity variation—the variation that cannot be explained by systematic effects. In Figure 9, we have predicted the individual observed slices using the Procrustes model and the proposed model. As we also saw in Figure 8, the proposed model predicts less deformation of the template compared to the Procrustes model, and we see that the Brownian sheet model is able to account for the majority of the personal structure in the sulci of the brain. Moreover, the predicted intensity variation seems to model intensity differences introduced by the different MRI scanners well.

6. Simulation study. In this section, we present a simulation study for investigating the precision of the proposed model. The results are compared to the previously introduced models: Procrustes free warp and a regularized Procrustes. Data are generated from model (2) in which θ is taken as one of the MRI slices considered in section 5.2. The warp, intensity, and the random noise effects are all drawn from the previously described multivariate normal

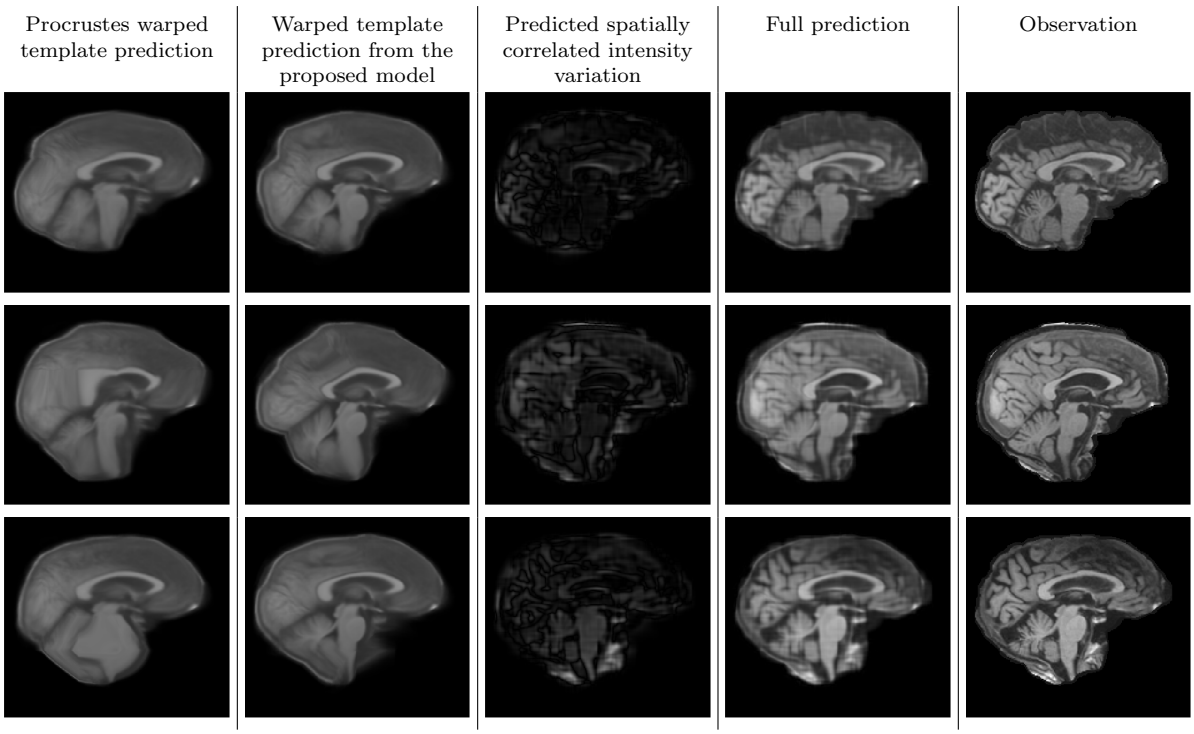

Figure 9. *Model predictions of three midsaggital slices (rightmost column). The first two rows display the warped templates from the Procrustes model and the proposed model. The third row displays the absolute value of the predicted spatially correlated intensity variation from the proposed model. The fourth row displays the full conditional prediction given the posterior warp variables* $\hat{\theta}(v(s,t,\hat{\boldsymbol{w}}_i)) + \mathrm{E}[x_i(s,t) \mid \boldsymbol{y}_i, \boldsymbol{w}_i = \hat{\boldsymbol{w}}_i]$.

distributions with variance parameters, respectively,

$$\sigma^2\gamma^2 = 0.01, \quad \sigma^2\tau^2 = 0.1, \quad \sigma^2 = 0.001,$$

and applied to the chosen template image θ. To consider more realistic brain simulations, the systematic part of the intensity effect was only added to the brain area of θ and not the background. As this choice makes the proposed model slightly misspecified, it will be hard to obtain precise estimates of the variance parameters. In practice, one would expect any model with a limited number of parameters to be somewhat misspecified in the presented setting. The simulations thus present a realistic setup and our main interest will be in estimating the template and predicting warp and intensity effects. Figure 10 displays 5 examples of the simulated observations as well as the chosen θ.

The study is based on 100 data sets of 100 simulated brains. For each simulated dataset we applied the proposed, Procrustes free warp, and Procrustes regularized model. The regularization parameter, λ, in the regularized Procrustes model, was set to the true parameter used for generating the data, $\lambda = \gamma^{-2}/2$.

The variance estimates based on the simulations are shown in Figure 11. The true variance parameters are plotted for comparison. We see some bias in the variance parameters. While bias is to be expected, the observed bias for the noise variance σ^2 and the warp variance scale

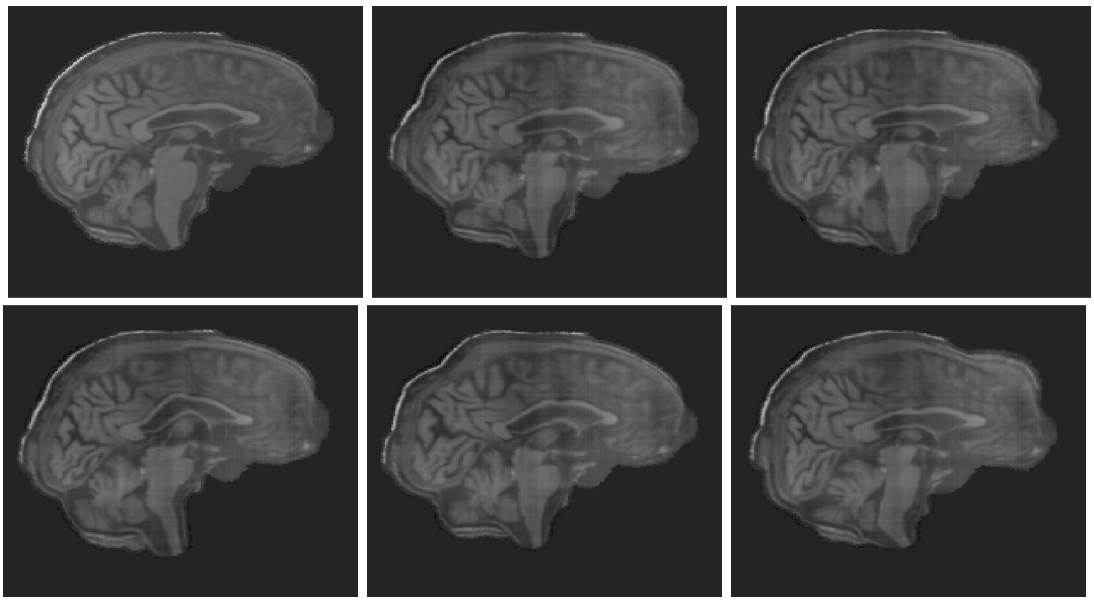

Figure 10. *Five examples of simulated brains. The template brain θ is shown in the upper left corner.*

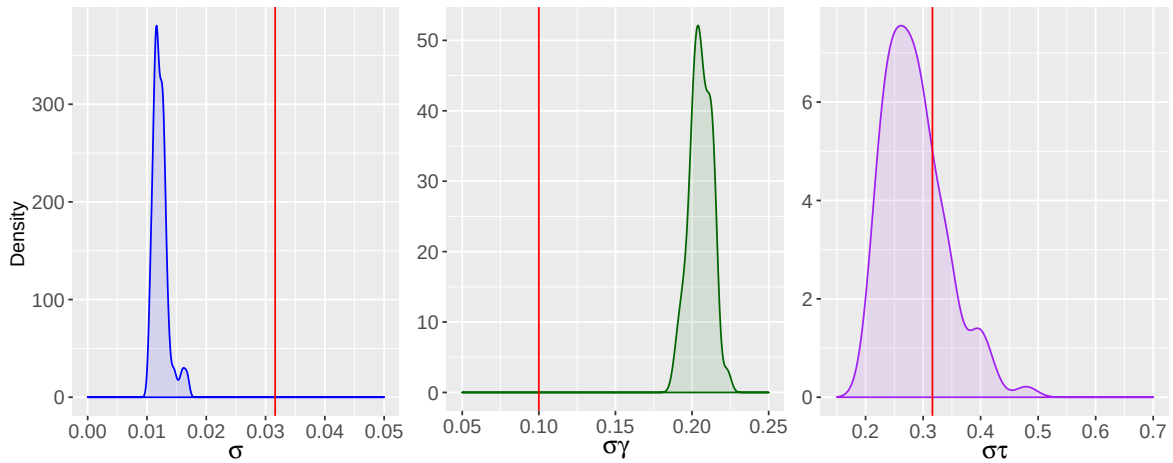

Figure 11. *Density plots for the estimated variance parameters in the proposed model. The vertical lines correspond to the true parameters.*

$\sigma^2\gamma^2$ are bigger than what one would expect. The reason for the underestimation of the noise variance seems to be the misspecification of the model. Since the model assumes spatially correlated noise outside of the brain area, where there is none, the likelihood assigns the majority of the variation in this area to the systematic intensity effect. The positive bias of the warp variance scale seems to be a compensating effect for the underestimated noise variance.

The left panel of Figure 12 shows the mean squared difference for the estimated templates θ with the three types of models. We see that the proposed model produces considerably more accurate estimates than the alternative frameworks.

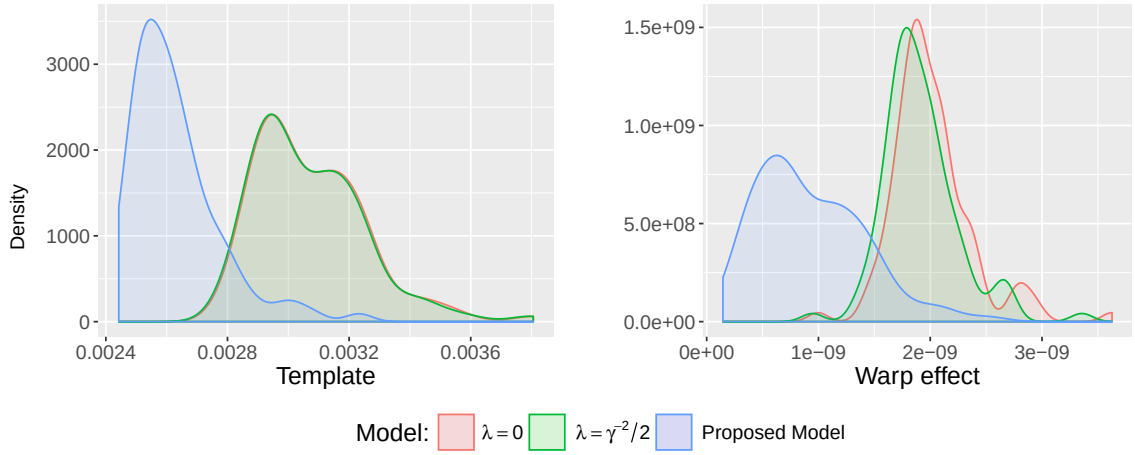

Figure 12. *Density plots for the mean squared differences of template and warp estimates for the three models. The plot to the left shows the density for the mean squared difference for the template effect and the plot to the right shows the mean squared difference for the warp effect.* $\lambda = 0$ *denotes the Procrustes free warp model,* $\lambda = \gamma^{-2}/2$ *is the Procrustes regularized model, and the blue density corresponds to the proposed model.*

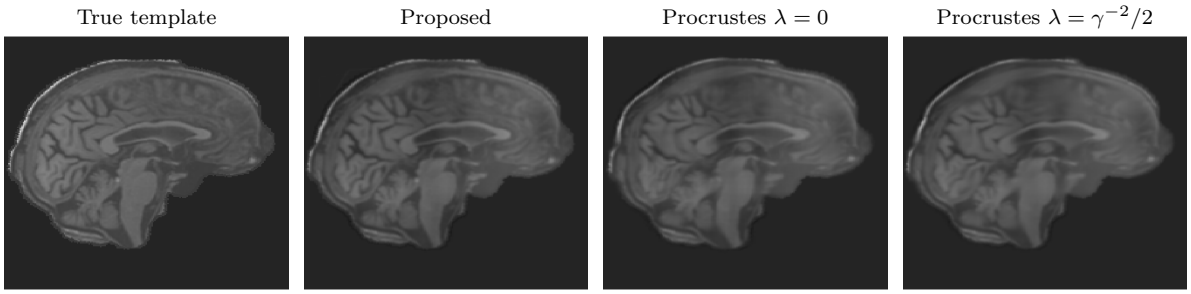

Figure 13. *Example of a template estimate for each of the three models. For comparison, the true* θ *are plotted as well.*

To give an example of the difference between template estimates for the three different models, one set of template estimates for each of the models is shown in Figure 13. From this example we see that the template for the proposed model is slightly more sharp than the Procrustes models and are more similar to the true θ which was also the conclusion obtained from the density of the mean squared difference for the template estimates (Figure 12).

The right panel of Figure 12 shows the mean squared prediction/estimation error of the warp effects. The error is calculated using only the warp effects in the brain area since the background is completely untextured, and any warp effect in this area will be completely determined by the prediction/estimation in the brain area. We find that the proposed model estimates warp effects that are closest to the true warps. It is worth noticing that the proposed model is considerably better at predicting the warp effects than the regularized Procrustes model. This happens despite the fact that the value for the warp regularization parameter in the model was chosen to be equal to the true parameter ($\lambda = \gamma^{-2}/2$). Examples of the true

warping functions in the simulated data and the predicted/estimated effects in the different models are shown in Figure 14. None of the considered models are able to make sensible predictions on the background of the brain, which is to be expected. In the brain region, the predicted warps for the proposed model seem to be very similar to the true warp effect, which we also saw in Figure 12 was a general tendency.

7. Conclusion and outlook. We generalized the likelihood-based mixed-effects model for template estimation and separation of phase and intensity variation from 1D images to 2D images. This type of model was originally proposed for curve data [34]. As the model is computationally demanding for high-dimensional data, we presented an approach for efficient likelihood calculations. We proposed an algorithm for doing maximum-likelihood-based inference in the model and applied it to two real-life datasets.

Based on the data examples, we showed how the estimated template had desirable properties and how the model was able to simultaneously separate sources of variation in a meaningful way. This feature eliminates the bias from conventional sequential methods that process data in several independent steps, and we demonstrated how this separation resulted in well-balanced trade-offs between the regularization of warping functions and intensity variation.

We made a simulation study to investigate the precision of the template and warp effects of the proposed model and for comparison with two other models. The proposed model was compared with a Procrustes free warp model, as well as a Procrustes regularized model. Since the noise model was misspecified, the proposed methodology could not recover precise maximum-likelihood estimates of the variance parameters. However, the maximum-likelihood estimate for the template was seen to be a lot sharper and closer to the true template compared to alternative Procrustes models. Furthermore, we demonstrated that the proposed model was better at predicting the warping effect than the alternative models.

The main restriction of the proposed model is the computability of the likelihood function. We resolved this by modeling intensity variation as a Gaussian Markov random field. An alternative approach would be to use the computationally efficient operator approximations of the likelihood function for image data suggested in [33]. This approach would, however, still require a specific choice of parametric family of covariance functions or, equivalently, a family of positive definite differential operators. An interesting and useful extension would be to allow a free low-rank spatial covariance structure and estimate it from the data. This could, for example, be done by extending the proposed model (2) to a factor analysis model where both the mean function and intensity variation are modeled in a common functional basis, and requiring a specific rank of the covariance of the intensity effect. Such a model could be fitted by means of an EM algorithm similar to the one for the reduced-rank model for computing functional principal component analysis proposed in [14], and it would allow simulation of realistic observations by sampling from the model.

For the computation of the likelihood function of the nonlinear model, we relied on local linearization which is a simple, well proven, and effective approach. In recent years, alternative frameworks for doing maximum-likelihood estimation in nonlinear mixed-effects models have emerged; see [6] and references therein. An interesting path for future work would be to formulate the proposed model in such a framework that promises better accuracy than the local linear approximation. This would allow one to investigate how much the linear ap-

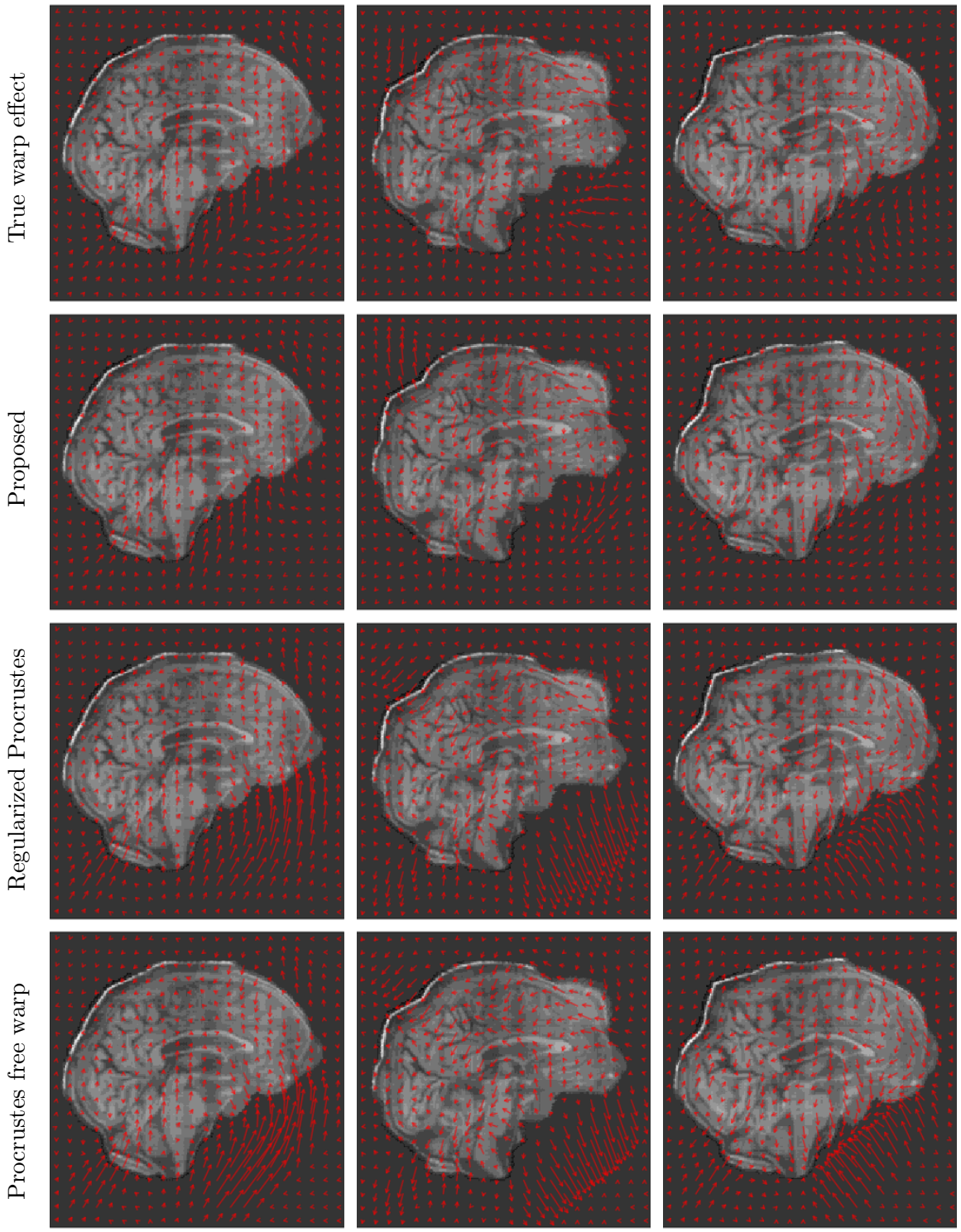

Figure 14. *Examples of predicted warp effect for each model. The top row shows the true warp effect, the second row the estimated warp effect of the proposed model, the third row regularized Procrustes, and the final row, the Procrustes model with free warps.*

proximation of the likelihood affects the estimated parameters. In this respect, it would also be interesting to compare the computing time across different methods to identify a suitable trade-off between accuracy and computing time.

The proposed model introduced in this paper is a tool for analyzing 2D images. The model, as it is, could be used for higher-dimensional images as well, but the analysis would be computationally infeasible with the current implementation. To extend the proposed model to 3D images there is a need to devise new computational methods for improving the calculation of the likelihood function.

REFERENCES

[1] S. Allassonnière, S. Durrleman, and E. Kuhn, *Bayesian mixed effect atlas estimation with a diffeomorphic deformation model*, SIAM J. Imaging Sci., 8 (2015), pp. 1367–1395.

[2] M. J. Black and P. Anandan, *The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields*, Comput. Vis. Image Underst., 63 (1996), pp. 75–104.

[3] V. Blanz and T. Vetter, *A morphable model for the synthesis of 3D faces*, in Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '99, ACM, New York, 1999, pp. 187–194.

[4] A. Bruhn, J. Weickert, and C. Schnörr, *Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods*, Int. J. Comput. Vis., 61 (2005), pp. 211–231.

[5] T. Cai, W. Liu, and X. Luo, *A constrained ℓ_1 minimization approach to sparse precision matrix estimation*, J. Amer. Statist. Assoc., 106 (2011), pp. 594–607.

[6] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell, *Stan: A probabilistic programming language*, J. Stat. Softw., 76 (2017).

[7] O. Demetz, D. Hafner, and J. Weickert, *The complete rank transform: A tool for accurate and morphologically invariant matching of structures*, in Proceedings of the 2013 British Machine Vision Conference, Bristol, UK, British Machine Vision Association and Society for Pattern Recognition, Durham, England, 2013.

[8] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9 (2008), pp. 432–441.

[9] R. Furrer and S. R. Sain, *spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields*, J. Stat. Softw., 36 (2010).

[10] D. Hafner, O. Demetz, and J. Weickert, *Why is the census transform good for robust optic flow computation?*, in International Conference on Scale Space and Variational Methods in Computer Vision, Springer, New York, 2013, pp. 210–221.

[11] D. Hafner, O. Demetz, J. Weickert, and M. Reissel, *Mathematical foundations and generalisations of the census transform for robust optic flow computation*, J. Math. Imaging Vision, 52 (2015), pp. 71–86.

[12] C. R. Henderson, *Estimation of genetic parameters*, Biometrics, 6 (1950), pp. 186–187.

[13] G. Hermosillo, C. Chefd'Hotel, and O. Faugeras, *Variational methods for multimodal image matching*, Int. J. Comput. Vis., 50 (2002), pp. 329–343.

[14] G. M. James, T. J. Hastie, and C. A. Sugar, *Principal component models for sparse functional data*, Biometrika, 87 (2000), pp. 587–602.

[15] A. Jorstad, D. Jacobs, and A. Trouve, *A deformation and lighting insensitive metric for face recognition based on dense correspondences*, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Piscataway, NJ, 2011, pp. 2353–2360.

[16] S. Joshi, B. Davis, B. M. Jomier, and G. Gerig B, *Unbiased diffeomorphic atlas construction for computational anatomy*, NeuroImage, 23 (2004), pp. 151–160.

[17] S. L. Lauritzen, *Graphical Models*, Oxford University Press, Oxford, 1996.

[18] M. J. Lindstrom and D. M. Bates, *Nonlinear mixed effects models for repeated measures data*, Biometrics, 46 (1990), pp. 673–687.

[19] J. Ma, M. I. Miller, A. Trouvé, and L. Younes, *Bayesian template estimation in computational anatomy*, NeuroImage, 42 (2008), pp. 252–261.
[20] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, *Multimodality image registration by maximization of mutual information*, IEEE Trans. Med. Imaging, 16 (1997), pp. 187–198.
[21] B. Markussen, *Functional data analysis in an operator-based mixed-model framework*, Bernoulli, 19 (2013), pp. 1–17.
[22] M. A. Mohamed and B. Mertsching, *TV-L1 optical flow estimation with image details recovering based on modified census transform*, in Advances in Visual Computing, Springer, Berlin, 2012, pp. 482–491.
[23] R. M. Neal and G. E. Hinton, *A view of the EM algorithm that justifies incremental, sparse, and other variants*, in Learning in Graphical Models, Kluwer, Dordrecht, Netherlands, 1998, pp. 355–368.
[24] S. Negahdaripour, *Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis*, IEEE Trans. Pattern Anal. Mach. Intell., 20 (1998), pp. 961–979.
[25] E. G. Ng and B. W. Peyton, *Block sparse Cholesky algorithms on advanced uniprocessor computers*, SIAM J. Sci. Comput., 14 (1993), pp. 1034–1056.
[26] A. Pai, S. Sommer, L. Sorensen, S. Darkner, J. Sporring, and M. Nielsen, *Kernel bundle diffeomorphic image registration using stationary velocity fields and Wendland basis functions*, IEEE Trans. Med. Imaging, 35 (2016), pp. 1369–1380.
[27] G. Panin, *Mutual information for multi-modal, discontinuity-preserving image registration*, in Advances in Visual Computing, Springer, Berlin, 2012, pp. 70–81.
[28] N. Papenberg, A. Bruhn, T. Brox, S. Didas, and J. Weickert, *Highly accurate optical flow computations with theoretically justified warping*, Int. J. Comput. Vis., 67 (2006), pp. 141–158.
[29] J. Pinheiro and D. Bates, *Mixed-Effects Models in S and S-PLUS*, Springer, New York, 2009.
[30] J. Pinheiro, D. Bates, S. DebRoy, and D. Sarkar, *Linear and Nonlinear Mixed Effects Models*, R Package version, 3 (2007), p. 57, https://CRAN.R-project.org/package=nlme.
[31] T. Pock, M. Urschler, C. Zach, R. Beichel, and H. Bischof, *A duality based algorithm for TV-L^1-optical-flow image registration*, in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2007, Springer, Berlin, 2007, pp. 511–518.
[32] L. L. Raket, *pavpop Version* 0.10, https://github.com/larslau/pavpop/ (2016).
[33] L. L. Raket and B. Markussen, *Approximate inference for spatial functional data on massively parallel processors*, Comput. Statist. Data Anal., 72 (2014), pp. 227–240.
[34] L. L. Raket, S. Sommer, and B. Markussen, *A nonlinear mixed-effects model for simultaneous smoothing and registration of functional data*, Pattern Recognit. Lett., 38 (2014), pp. 1–7.
[35] G. K. Robinson, *That BLUP is a good thing: The estimation of random effects*, Statist. Sci., 6 (1991), pp. 15–32.
[36] A. Roche, G. Malandain, X. Pennec, and N. Ayache, *The correlation ratio as a new similarity measure for multimodal image registration*, in Medical Image Computing and Computer-Assisted Interventation–MICCAI–98, Springer, Berlin, 1998, pp. 1115–1124.
[37] H. Rue and H. Tjelmeland, *Fitting Gaussian Markov random fields to Gaussian fields*, Scand. J. Stat., 29 (2002), pp. 31–49.
[38] F. S. Samaria, *The Database of Faces*, http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html.
[39] A. Sotiras, C. Davatzikos, and N. Paragios, *Deformable medical image registration: A survey*, IEEE Trans. Med. Imaging, 32 (2013), pp. 1153–1190.
[40] L. Su, A. M. Blamire, R. Watson, J. He, B. Aribisala, and J. T. O'Brien, *Cortical and subcortical changes in Alzheimer's disease: A longitudinal and quantitative MRI study*, Curr. Alzheimer Res., 13 (2016), pp. 534–544.
[41] A. Trouvé and L. Younes, *Local geometry of deformable templates*, SIAM J. Math. Anal., 37 (2005), pp. 17–59.
[42] A. Trouvé and L. Younes, *Metamorphoses through Lie group action*, Found. Comput. Math., 5 (2005), pp. 173–198.
[43] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, *N4ITK: Improved N3 bias correction*, IEEE Trans. Med. Imaging, 29 (2010), pp. 1310–1320.
[44] R. Wolfinger, *Laplace's approximation for nonlinear mixed models*, Biometrika, 80 (1993), pp. 791–795.

[45] X. Xie and K.-M. Lam, *Face recognition using elastic local reconstruction based on a single face image*, Pattern Recognit., 41 (2008), pp. 406–417.
[46] L. Younes, *Shapes and Diffeomorphisms*, Springer, Heidelberg, 2010.
[47] R. Zabih and J. Woodfill, *Non-parametric local transforms for computing visual correspondence*, in Computer Vision—ECCV'94, Springer, Berlin, 1994, pp. 151–158.
[48] M. Zhang, N. Singh, and P. T. Fletcher, *Bayesian estimation of regularization and atlas building in diffeomorphic image registration*, in Information Processing for Medical Imaging (IPMI), Lecture Notes in Comput. Sci. 7917, Springer, Berlin, 2013, pp. 37–48.