Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

*Proc IEEE Int Symp Biomed Imaging*. Author manuscript; available in PMC 2018 December 11.

Published in final edited form as:

*Proc IEEE Int Symp Biomed Imaging*. 2017 April ; 2017: 868–872. doi:10.1109/ISBI.2017.7950654.

# ADAPTIVE GRADIENT DESCENT OPTIMIZATION OF INITIAL MOMENTA FOR GEODESIC SHOOTING IN DIFFEOMORPHISMS

**Greg M. Fleishman** and

UCLA Bioengineering, 420 Westwood Plaza, 5121 Engineering V, UCLA, CA 90095-1600

**Paul M. Thompson**

USC, Imaging Genetics Center, 4676 Admiralty Way, 2nd floor, Marina del Rey, CA 90292

## Abstract

Diffeomorphic image registration algorithms are widely used in medical imaging, and require optimization of a high-dimensional nonlinear objective function. The function being optimized has many characteristics that are relevant for optimization but are typically not well understood. Due to that complexity, most authors have used a simple gradient descent, but it is not often discussed how step sizes are chosen or if line searches are used. Further, if a system is to be robust to a range of input images, that may differ to varying degrees, the optimization must be adaptable. Here, we present two methods of adaptable gradient descent with line searches, and test how they affect image registration. The optimization schemes are deployed for geodesic shooting in diffeomorphisms - an approach that is used to quantify anatomical changes, such as atrophy, in longitudinal image pairs. We evaluate the optimization schemes on their convergence characteristics and based on how well the resulting atrophy scores correlate with diagnostic group and mini mental state exam (MMSE) scores. We find that the Barzilai-Borwein method with a backtracking line search outperforms other optimization schemes in convergence time and adaptability by a wide margin. We also find that the variable optimization schemes do not significantly affect the ability to measure atrophy with clinical significance.

### Index Terms

LDDMM; diffeomorphisms; geodesic shooting; gradient descent; optimization[1]

## 1. INTRODUCTION

The LDDMM (large deformation diffeomorphic metric mapping) framework for diffeomorphic image registration is described in significant procedural detail in the seminal paper by Beg et al. [1]. LDDMM proposes encoding the shape difference evident in two images of the same anatomy as a point on a manifold of diffeomorphisms. The objective of the algorithm is to construct a path on that manifold beginning at the identity and ending at the diffeomorphism that optimally matches the two images; Beg et al. show that at optimality the path is a geodesic. In the geodesic shooting formulation, the path is parameterized by an initial momentum vector field, from which the entire geodesic can be reconstructed by integrating the appropriate Euler-Poincaré differential equation (EPdiff) [2, 3, 4, 5, 6]. Also, Miller et al. [2] show that at optimality, the initial momentum vector field is

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

proportional to the spatial gradient of the moving image. Hence, the objective of geodesic shooting in diffeomorphisms (GSiD) is to find a scalar momentum field that parameterizes an optimal matching between the given images.

An implementation of GSiD can be viewed as having three mathematical components: (1) construction of the model itself, which can include decisions about representation [6] and selection of parameters that dictate properties of the space of diffeomorphisms [7], (2) numerical integration of differential equations [3], and (3) an optimization procedure for the initial momentum field. The majority of work in the field has been in areas (1) and (2), with substantially less attention to paid to (3). With the exception of [5], most studies report using gradient descent with some step size $\varepsilon$, though details of the procedure including the determination of $\varepsilon$, are typically omitted. Very little information relevant to optimization is known about the GSiD objective function, such as its smoothness and curvature characteristics. Further, it is not known how variable these characteristics are to different inputs. Nonetheless, to build a GSiD system robust to variable inputs, some optimization procedure must be selected.

We consider here gradient descent with three different procedures to determine the step size $\varepsilon$: (1) a static step size, (2) a secant method line search, and (3) the Barzilai-Borwein method [8]. Methods (2) and (3) compute the step size at every iteration using limited local curvature data estimated from the objective function; hence, $\varepsilon$ is adaptable to the inputs and the particular iteration of the optimization.

## 2. METHODS

### 2.1. GSiD

A complete discussion of the GSiD model is beyond the scope of this paper; for a thorough discussion of the following equations see [3]. For a moving image $I$ and fixed image $J$, the GSiD objective function is:

$$E(P_0) = \frac{1}{\sigma^2}\langle P_0 \nabla I, K(P_0 \nabla I)\rangle_{L_2} + \left\| I \circ \phi_1^{-1} - J \right\|^2 \quad (1)$$

which must be minimized with respect to the initial scalar momentum field $P_0$. A given initial momentum provides the initial conditions for the EPdiff equation(s), which govern the time evolution of the momentum and moving image:

$$\begin{cases} \partial_t I + \nabla I \cdot v = 0 \\ \partial_t P + \nabla \cdot (P v) = 0 \\ v + K(P \nabla I) = 0 \end{cases} \quad (2)$$

The third equation states the relationship between momentum and velocity, where $K$ plays the role of an inertia; $K$ is a smoothing kernel and $K(w)$ is taken to mean the convolution of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

vector field $w$ with $K$. The path of diffeomorphisms $\phi_t$ is constructed from $v(x, t)$ according to the ODE:

$$\begin{cases} \partial_t \phi_t = v_t(\phi) \\ \phi_0 = \mathrm{Id} \end{cases} \quad (3)$$

This yields the geodesic path of diffeomorphisms $\phi_t$, where the end point $\phi_1$ is used to match $I$ and $J$. The matching residual $J - I \circ \phi_1^{-1}$, which resides in the coordinate system of $J$, must be brought back to the coordinate system of $I$ while respecting the geodesicity of the whole path $\phi_t$. This is done by integrating the adjoint system backwards in time with initial conditions $\hat{I}_1 = J - I \circ \phi_1^{-1}$ and $\hat{P}_0 = 0$:

$$\begin{cases} \partial_t \hat{I} + \nabla \cdot \left(v\hat{I}\right) + \nabla \cdot (P\hat{v}) = 0 \\ \partial_t \hat{P} + v \cdot \nabla \hat{P} + \nabla I \cdot \hat{v} = 0 \\ \hat{v} + K\left(\hat{I} \nabla I - P \nabla \hat{P}\right) = 0 \end{cases} \quad (4)$$

$\hat{P}_0$ completes the gradient of equation (1) with respect to $P_0$. In a gradient descent scheme, that gradient is used to update $P_0$:

$$P_0^{k+1} = P_0^k - \varepsilon\left(\nabla I \cdot K\left(P_0^k \nabla I\right) - \hat{P}_0^k\right) \quad (5)$$

$\varepsilon$ is one of the few user selected parameters in the GSiD model. A poor selection of $\varepsilon$ can result in intractable compute times (if the user insists on running to convergence), sub-convergent results (if the optimization is stopped early due to time considerations), or numerical instability and divergence (if $\varepsilon$ is too large). We are concerned in particular with the application of GSiD to longitudinal MRI time series of the brain to quantify atrophy. In that application, deformations are typically very low amplitude even relative to the spatial resolution of the images. Despite that, atrophy of as little as 5% in critical brain areas can have a significant impact on quality of life [9]. Hence, it is crucial to measure longitudinal deformations with the highest degrees of accuracy and precision possible. In such a case, the selection of $\varepsilon$ can be critical to ensuring accurate and unbiased measurements.

### 2.2. Adaptable gradient descent steps

The simplest option is to select *a priori* a static value for $\varepsilon$ which is fixed throughout the optimization. This value may perform well for some instances of data, and poorly for others. Even for a fixed input, it may perform well for a subset of iterations and poorly for others. We include this option as a baseline for comparison with more intelligent choices.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**The secant line search method—**Suppose we would like to optimize a function $f(x)$ by gradient descent. Then, for iteration $k$, we would like to minimize $f(x_k - \varepsilon_k f'(x_k))$ with respect to $\varepsilon_k$. Take the truncated Taylor expansion (we temporarily omit the subscript $k$):

$$f(x-\varepsilon f') \approx f(x) + \varepsilon\left(\partial_\varepsilon f(x-\varepsilon f')\big|_{\varepsilon=0}\right) + \frac{\varepsilon^2}{2}\left(\partial_\varepsilon^2 f(x-\varepsilon f')\big|_{\varepsilon=0}\right) \quad (6)$$

If used directly, the second-order term will require the Hessian matrix of $f$. GSiD is a very high-dimensional optimization, hence the Hessian matrix $f''$ is intractable. We can replace the second-order term in the Taylor expansion with a finite difference approximation on the gradients:

$$\begin{aligned}\partial_\varepsilon^2 f(x-\varepsilon f') &\approx \frac{\partial_\varepsilon f(x-\varepsilon f')\big|_{\varepsilon=\sigma} - \partial_\varepsilon f(x-\varepsilon f')\big|_{\varepsilon=0}}{\sigma} \quad (7)\\ &= \frac{-f'(x-\sigma f')^T f' + f'^T f'}{\sigma}\end{aligned}$$

Substitute (7) into (6), apply the remaining partial derivative, and further differentiate each side with respect to $\varepsilon$. You will arrive at the expression:

$$\partial_\varepsilon f(x-\varepsilon f') \approx -f'^T f' + \frac{\varepsilon}{\sigma}\left(f'^T f' - f'(x-\sigma f')^T f'\right)$$

Finally, we set this equal to zero and solve for $\varepsilon$. Also, to use this formula for GSiD we must account for the metric in the space of momenta; the inner products must include the operator $K$. With these two final steps we arrive at the formula:

$$\varepsilon_k = \frac{\sigma_k f'^T_k K(f'_k)}{f'^T_k K(f'_k) - f'(x_k - \sigma_k f')^T K(f'_k)} \quad (8)$$

Where $f'_k = f'(x_k)$. Essentially, the secant method approximates the objective function in the gradient direction as a parabola, the curvature of which is estimated by formula (7). Because the function may not be well estimated locally as a parabola, for a given gradient descent iteration $k$, formula (8) is applied iteratively giving a series of steps $\varepsilon_k^i$. For the $i^{th}$ secant method iteration, $\sigma_k^i = -\varepsilon_k^{i-1}$, which leverages every gradient computation efficiently. $\sigma_k^0$ is set to a default value. This line search is stopped after a certain number of fixed iterations or when the magnitude of the update $\|\varepsilon f'_k\|$ falls below a threshold. Note, though we must evaluate multiple gradients during the line search iterations, we only move in the direction $f'_k$ until the line search is stopped and we move to gradient descent iteration $k+1$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

**The Barzilai-Borwein method—**We derive the method assuming $f(x) = \frac{1}{2}x^T Ax - b^T x$. The gradient is then $f' = Ax - b$ and the Hessian is $f'' = A$. Newton's method, a second-order optimization that accounts for the curvature of the objective function, proceeds as $x_{k+1} = x_k - A^{-1} f'_k$. (For a symmetric positive definite quadratic form, this will converge in a single step and is equivalent to Gaussian elimination). The objective of the Barzilai-Borwein (BB) method is to let $\varepsilon$ be determined by the simplest possible approximation to Newton's method:

$$-\varepsilon f' = -\left(\varepsilon^{-1}\mathrm{Id}\right)^{-1} f' \approx -A^{-1} f' \quad (9)$$

Let $s_k = x_k - x_{k-1}$ and $y_k = f'_k - f'_{k-1}$. For the quadratic form, $A$ satisfies $As_k = y_k$. So, we will let $\varepsilon$ be the solution to the least squares problem:

$$\varepsilon_k = \mathrm{argmin}_\alpha \frac{1}{2}\left\| s_k - \alpha y_k \right\|^2 \quad (10)$$

which has the closed-form solution:

$$\varepsilon_k = \frac{s_k^T y_k}{y_k^T y_k} \quad (11)$$

Again, to apply this to GSiD we must account for the metric in the space of momenta:

$$\varepsilon_k = \frac{s_k^T K(y_k)}{y_k^T K(y_k)} \quad (12)$$

Similar to the secant method, the BB method approximates second-order information, but it does not require a second gradient computation. Even so, in some places formula (9) is likely to be a very poor approximation. It is well known that as a result, BB step sizes do not provide monotonic optimization; that is, occasionally $\varepsilon_k$ is too large. However, for nonlinear optimization, some degree of nonmonotonicity may be desired as it may help escape spurious local minima. Hence, the BB method is often coupled with a backtracking line search [10]. In our case, $\varepsilon_k$ is iteratively cut in half until the first Wolfe condition is satisfied:

$$f\left(x_k - \varepsilon_k f'_k\right) \leq \max_j f\left(x_j\right) - \gamma \varepsilon_k f'^T_k K\left(f'_k\right) \quad (13)$$

where $\max_{(k-M,0)} \leq j \leq k$. $M$ controls the degree of monotonicity (we use $M$=10) and $\gamma$ is related to our expectation of the objective function's local curvature. $\gamma$ is typically chosen to be small (we use $\gamma = 10^{-4}$).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

## 3. EXPERIMENTAL RESULTS

We took 100 randomly chosen subjects from the ADNI-2 longitudinal MRI dataset - available at adni.loni.usc.edu - and registered their baseline scans to their 24-month follow-up scans using GSiD. We did these registrations under 5 different experimental conditions: static step sizes of 0.001, 0.01, and 0.1, the secant method with $\varepsilon_k^0 = 0.01$, and the BB method with $\varepsilon_0 = 0.01$. For all approaches, the optimization was stopped when the gradient magnitude (relative to the initial gradient magnitude) fell below a chosen threshold, or after 300 iterations, whichever came first. Before GSiD, the images were preprocessed according to the protocol detailed in [11, 12].

After GSiD, the Jacobian determinants of the deformations mapping the baseline to the 24 month followup images were moved to a common coordinate system. The Jacobian determinants were averaged in a region where the rate of atrophy is significantly associated with Alzheimer's Disease (AD), a stat-ROI, (Fig. 1) to produce a scalar value atrophy score that represents the percent volume loss within the region for each subject [9]. The region was constructed from a non-overlapping data set from that evaluated here.

Figures 2 and 3 show convergence characteristics of the five optimizations strategies. Contrary to equation (1), we did not use sum of squared differences to drive the registration. We used the squared Local Correlation Coefficient (LCC) which is also used in [13]; LCC increases as the images become better matched. Curves that do not extend the full 300 iterations are instances that stopped early due to the gradient magnitude stopping criterion. The largest static step size clearly causes oscillations in all instances. The smallest static step size did not permit any instances to complete before reaching 300 iterations, it is likely that many instances are sub-convergent. The middle static step size appears to be a good compromise, but for many instances, the gradient magnitude oscillates. Apparently, none of the static step sizes is appropriate for all instances of the data or through all iterations of the optimization. The secant and BB methods show better convergence characteristics, with more instances finishing early. However, for the secant method not all instances converged. The spikes in the gradient magnitude for the BB method are due to the nonmonotonicity discussed above.

A good first question to ask is whether the choice of optimization procedure had a significant impact on atrophy scores. Figure 4 shows the p-values from paired t-tests between the measured atrophy scores for all pairs of optimization approaches. All five optimization procedures produced atrophy measurements that were significantly different from the others. Figure 3 also shows the average number of iterations and the number of instances that failed to converge due to numerical instability. In practice, the failed instances would have to be rerun with the parameters adjusted by hand. The BB method clearly had the fastest convergence, and was sufficiently adaptable that no instances failed to converge.

Atrophy measurements such as these have been shown to correlate with diagnostic category and performance on cognitive tests. The data set included subjects from four diagnostic categories: healthy controls (HC), early mild cognitive impairment (eMCI), late mild cognitive impairment (lMCI), and Alzheimer's disease (AD). Each subject also had a mini

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

mental state exam (MMSE) administered at the 24 month follow up time point. The MMSE scores from 0 - 30, where scores below 24 typically indicate some level of dementia. Figure 4 also shows Pearson's correlation coefficients between atrophy scores and diagnostic group and also between atrophy scores and MMSE scores for each of the five optimization approaches. The correlations appear sufficiently similar across optimization approaches to suggest that faster or more adaptable optimization approaches do not compromise the ability to measure clinically meaningful atrophy.

## Acknowledgments

The authors would like to thank Dr. Christian Clason for a helpful discussion on nonlinear optimization. This work was supported, in part, by NIH grant U54 EB020403 to the ENIGMA Center for Worldwide Medicine, Imaging Genomics.

## References

1. Beg M FaisalMiller Michael I, Trouvé AlainYounes Laurent. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. Int J Comput Vision. Feb; 2005 61(2):139–157.
2. Miller Michael I, Trouve AlainYounes Laurent. Geodesic shooting for computational anatomy. Journal of Mathematical Imaging and Vision. 2006; 24(2):209–228. [PubMed: 20613972]
3. Vialard Francois-XavierRisser LaurentRueckert DanielCotter Colin J. Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation. Int J Comput Vision. Apr; 2012 97(2):229–241.
4. Niethammer MarcHuang YangVialard Francois-Xavier. MICCAI. 2011, vol 6892 of Lecture Notes in Computer Science. Springer; Geodesic regression for image time-series; 655–662.
5. Ashburner JohnFriston Karl J. Diffeomorphic registration using geodesic shooting and GaussNewton optimisation. NeuroImage. 2011; 55(3):954–967. [PubMed: 21216294]
6. Singh NP, Hinkle J, Joshi S, Fletcher PT. A vector momenta formulation of diffeomorphisms for improved geodesic regression and atlas construction. Proceedings of the 2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI). 2013:1219–1222.
7. Risser LaurentVialard Francois-Xavier X, Wolz RobinHolm Darryl D, Rueckert Daniel. Simultaneous fine and coarse diffeomorphic registration: application to atrophy measurement in Alzheimer's disease. MICCAI. 2010; 13(Pt 2):610–617. [PubMed: 20879366]
8. Barzilai JonathanBorwein Jonathan M. Two-Point Step Size Gradient Methods. IMA Journal of Numerical Analysis. Jan; 1988 8(1):141–148.
9. Hua XueChing Christopher RK, Mezher AdamGutman BorisHibar Derrek P, Bhatt PriyaLeow Alex D, Jack Clifford R, , JrBernstein MattWeiner Michael W, Thompson Paul M. MRI-based brain atrophy rates in ADNI phase 2: acceleration and enrichment considerations for clinical trials. Neurobiology of Aging. 2016; 37:26–37. [PubMed: 26545631]
10. Fletcher Roger. On the Barzilai-Borwein Method. Springer US; 2005. 235–256.
11. Fleishman Greg M, Gutman Boris A, Fletcher P ThomasThompson Paul M. Simultaneous longitudinal registration with group-wise similarity prior. International Conference on Information Processing in Medical Imaging Springer International Publishing. 2015:746–757.
12. Fleishman Greg M, Fletcher P ThomasGutman Boris A, Prasad GautamWu YingnianThompson Paul M. Geodesic refinement using james-stein estimators. Mathematical Foundations of Computational Anatomy, MICCAI. 2015:60.
13. Avants Brian B, Yushkevich Paul A, Pluta JohnMinkoff DavidKorczykowski MarcDetre John A, Gee James C. The optimal template effect in hippocampus studies of diseased populations. NeuroImage. 2010; 49(3):2457–2466. [PubMed: 19818860]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

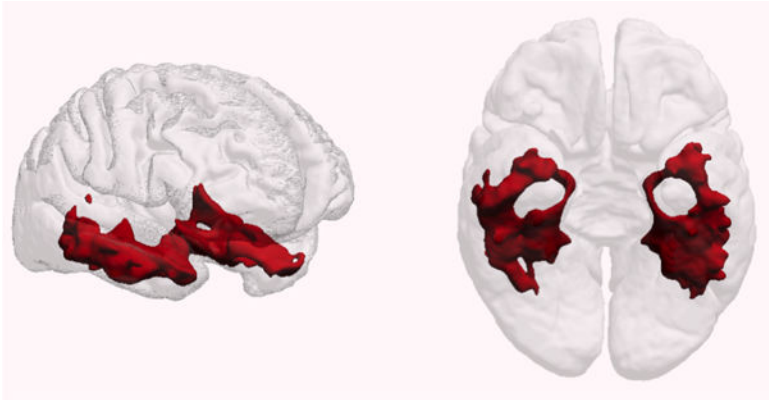

**Fig. 1.**
Region of interest with significant atrophy in AD, used here to compute atrophy scores

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

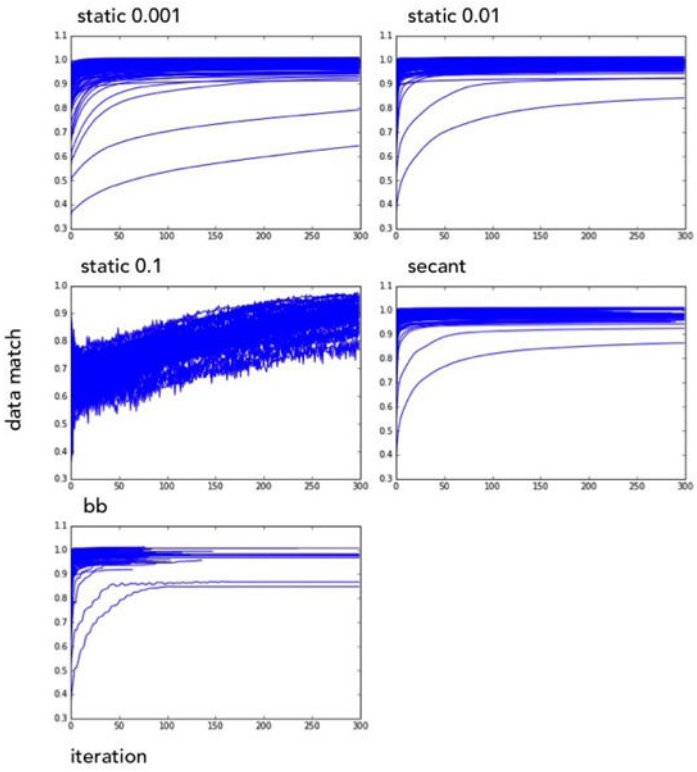


**Fig. 2.**

Convergence of data match term; curves that do not extend the full 300 iterations stopped early due to the gradient magnitude stopping criteria.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

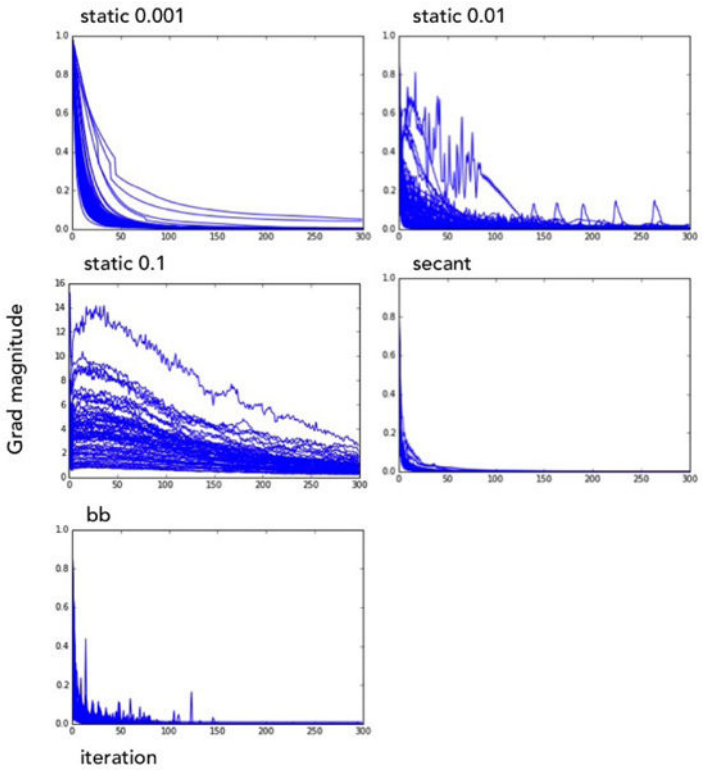


**Fig. 3.**

Convergence of gradient magnitude; curves that do not extend the full 300 iterations stopped early due to the gradient magnitude stopping criteria.

| paired t-test p vals | Static 0.001 | Static 0.01 | Static 0.1 | Secant | BB |
|---|---|---|---|---|---|
| Static 0.001 | | 1.92E-11 | 2.59E-12 | 1.67E-10 | 2.51E-12 |
| Static 0.01 | | | 3.33E-16 | 7.69E-04 | 3.24E-03 |
| Static 0.1 | | | | 4.11E-15 | 8.88E-16 |
| Secant | | | | | 3.49E-02 |
| BB | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Average # of iterations | 300 | 248 | 300 | 227 | 80 |
| # that diverged | 0 | 0 | 39 | 17 | 0 |
| DX Correlation | 0.534 | 0.521 | 0.548 | 0.467 | 0.519 |
| MMSE Correlation | -0.732 | -0.7 | -0.697 | -0.694 | -0.709 |

**Fig. 4.**

Statistical tests, convergence information, and correlations; DX: diagnostic group; MMSE: Mini Mental State Exam



Author Manuscript Author Manuscript Author Manuscript Author Manuscript