Incorporating Prior Knowledge Into Image Registration.

J Ashburner[1] P Neelin[2] D L Collins[2] A Evans[2] K Friston[1]

1. Wellcome Department of Cognitive Neurology,
   Institute of Neurology, London, UK
2. McConnell Brain Imaging Centre,
   Montreal Neurological Institute,
   Montreal, Quebec, Canada

Running Title Bayesian Image Registration

Keywords affine, Bayesian, MAP, registration, PET, MRI.

Address for correspondence
Wellcome Department of Cognitive Neurology,
Functional Imaging Laboratory,
12 Queen Square,
London. WC1N 3BG
U.K.

Tel +44 (0)171 833 7472

Fax +44 (0)171 813 1420

email j.ashburner@fil.ion.ucl.ac.uk

Abstract

The first step in the spatial normalization of brain images, is usually to determine the affine transformation that best maps the image to a template image in a standard space. We have developed a rapid and automatic method for performing this registration, which uses a Bayesian scheme to incorporate prior knowledge of the variability in the shape and size of heads. We compared affine registrations with and without incorporating the prior knowledge. We found that the affine transformations derived using the Bayesian scheme are much more robust, and that the rate of convergence is greater.

1 Introduction.

In order to average signals from functional brain images of different subjects, it is necessary to register the images together. This is often done by mapping all the images into the same standard space (Talairach & Tournoux, 1988). Almost all between subject co-registration or spatial normalization methods for brain images begin with determining the optimal 9 or 12 parameter affine transformation that registers the images together. This step is normally performed automatically by minimizing (or maximizing) some mutual function of the images. Without constraints and with poor data, the simple parameter optimization approach can produce some extremely unlikely transformations. For example, when there are only a few transverse slices in the image (spanning the X and Y dimensions), it is not possible for the algorithms to determine an accurate zoom in the Z direction. Any estimate of this value is likely to have very large errors. Previously in this situation, it was better to assign a fixed value for this difficult-to-determine parameter, and simply fit for the remaining ones.

By incorporating prior information into the optimization procedure, a smooth transition between fixed and fitted parameters can be achieved. When the error for a particular fitted parameter is known to be large, then that parameter will be based more upon the prior information. The approach adopted here is essentially a *maximum a posteriori* (*MAP*) Bayesian approach.

The *Methods* section of this paper begins by explaining the basic optimization method that is used, before introducing the principles behind the Bayesian approach. This is followed by sections on how the errors in the parameter increments are determined, and how the *a priori* probability distributions were derived. The modifications to the basic iterative scheme in order to incorporate the prior information are then presented.

The *Results and Discussion* section illustrates the potential benefit of the Bayesian approach, by showing both faster convergence for good data and improved parameter estimates for limited data. The paper ends with a discussion on the implications of a Bayesian approach for non-linear image registration.

2 Methods.

2.1 The Basic Optimization Method.

The objective is to fit the image $\mathbf{f}$ to a template image $\mathbf{g}$, using a twelve parameter affine transformation (parameters p_1 to p_{12}). The images may be scaled quite differently, so we also need to include an additional intensity scaling parameter (p_{13}) in the model.

An affine transformation mapping (via matrix $\mathbf{M}$, where the matrix elements are a function of parameters p_1 to p_{12}) from position $\mathbf{x}$ in one image to position $\mathbf{y}$ in another is defined by:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ 1 \end{pmatrix} = \begin{pmatrix} m_{1,1} & m_{1,2} & m_{1,3} & m_{1,4} \\ m_{2,1} & m_{2,2} & m_{2,3} & m_{2,4} \\ m_{3,1} & m_{3,2} & m_{3,3} & m_{3,4} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}$$

We refer to this mapping as $\mathbf{y} = \mathbf{M}\mathbf{x}$.

The parameters ($\mathbf{p}$) are optimized by minimizing the sum of squared differences between the images according to the *Gauss Newton* algorithm as described in Friston *et al.*(1995b). The function that is minimized is:

$$\sum_{i=1}^{I} \left(f(\mathbf{M}\mathbf{x}_i) - p_{13} g(\mathbf{x}_i)\right)^2$$

The optimization method involves generating a linear approximation to the problem using Taylor's Theorem, which is solved on each iteration (see Press *et al.*(1992), Section 15.5 for a full explanation of the approach). For iteration n, this can be expressed as computing:

$$\mathbf{p}^{(n)} = \mathbf{p}^{(n-1)} - \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\left(\mathbf{A}^T\mathbf{b}\right) \tag{1}$$

where element b_i of $\mathbf{b}$ is the ith residual $(f(\mathbf{M}\mathbf{x}_i) - p_{13} g(\mathbf{x}_i))$ and element $a_{i,j}$ of the Jacobian matrix $\mathbf{A}$ is the derivative of residual b_i with respect to parameter p_j. The approximation is most valid when the estimates are close to the true solution, relative to the smoothness of the image. Because of this, the images are smoothed prior to matching.

The rate of change of the residual b_i with respect to the scaling parameter (p_{13}) is simply $-g(\mathbf{x}_i)$ (the negative intensity of image $\mathbf{g}$ at $\mathbf{x}_i$ - the ith sample position).

The derivatives of the residuals with respect to the spatial transformation parameters (p_1 to p_{12}) are obtained by differentiating $f(\mathbf{M}\mathbf{x}_i) - p_{13}g(\mathbf{x}_i)$ with respect to p_j to give:

$$a_{ij} = \frac{\partial f(\mathbf{M}\mathbf{x}_i)}{\partial p_j}$$

There are many ways of parameterizing an affine transformation. The simplest parameters to optimize are the elements of the transformation matrix. The ith derivative of the residuals with respect to changes in element $m_{j,k}$ of matrix $\mathbf{M}$ is $x_{k,i}\frac{\partial f(\mathbf{y})}{\partial y_j}$ for elements $m_{1,1}$ to $m_{3,3}$, and simply $\frac{\partial f(\mathbf{y})}{\partial y_j}$ for elements $m_{1,4}$ to $m_{3,4}$, where $\mathbf{y} \equiv \mathbf{M}\mathbf{x}_i$.

The optimization can however be easily re-parameterized from parameter set $\mathbf{p}$ to parameter set $\mathbf{q}$, simply by incorporating an additional matrix $\mathbf{R}$ such that $r_{i,j} = dp_j/dq_i$. This matrix is re-computed in each iteration. The iterative scheme would then become $\mathbf{q}^{(n)} = \mathbf{q}^{(n-1)} - (\mathbf{R}^T(\mathbf{A}^T\mathbf{A})\mathbf{R})^{-1}\mathbf{R}(\mathbf{A}^T\mathbf{b})$ (the braces indicate the most efficient way of performing the computations). Extensions of the approach described in this paper require this re-parameterization, but for simplicity it will not be included in the description of the basic method.

In this implementation, the distance between samples is every eight millimeters (rounded to the nearest whole number of voxels in image $\mathbf{g}$). Tri-linear interpolation of the voxel lattice (rather than the sampling lattice) is used to resample the images at the desired co-ordinates. Gradients of the images are obtained at the same time, using a finite difference method on the same voxel lattice. No assumptions are made about voxel values that lie outside the field of view of image $\mathbf{f}$. Points where $\mathbf{M}\mathbf{x}_i$ falls outside the domain of $\mathbf{f}$ are not included in the computations.

2.2 A Bayesian Approach.

Bayes rule is generally expressed in the continuous form:

$$p(a_{\mathbf{p}}|b) = \frac{p(b|a_{\mathbf{p}})p(a_{\mathbf{p}})}{\int_{\mathbf{q}} p(b|a_{\mathbf{q}})p(a_{\mathbf{q}})d\mathbf{q}}$$

where $p(a_{\mathbf{p}})$ is the prior probability of $a_{\mathbf{p}}$ being true, $p(b|a_{\mathbf{p}})$ is the conditional probability that b is observed given that $a_{\mathbf{p}}$ is true and $p(a_{\mathbf{p}}|b)$ is the Bayesian estimate of $a_{\mathbf{p}}$ being true, given that measurement b has been made. The expression $\int_{\mathbf{q}} p(b|a_{\mathbf{q}})p(a_{\mathbf{q}})d\mathbf{q}$ is included so that the total probability of all possible outcomes is unity. The *maximum a posteriori* estimate for parameters $\mathbf{p}$ is the mode of $p(a_{\mathbf{p}}|b)$. For our purposes, $p(a_{\mathbf{p}})$ represents a known prior probability distribution from which the parameters are drawn, $p(b|a_{\mathbf{p}})$ is the likelihood of obtaining the parameters given the data b and $p(a_{\mathbf{p}}|b)$ is the function to be maximized. The optimization can be simplified by assuming that all probability distributions are multidimensional and normal (multi-normal), and can therefore be described by a mean vector and a covariance matrix.

When close to the minimum, the optimization becomes almost a linear problem. This allows us to assume that the errors of the fitted parameters ($\mathbf{p}$) can be locally approximated by a multi-normal distribution with covariance matrix $\mathbf{C}$. We assume that the true parameters are drawn from an underlying multi-normal distribution of known mean ($\mathbf{p_0}$) and covariance ($\mathbf{C_0}$). By using the *a priori* probability density function (p.d.f) of the parameters, we can obtain a better estimate of the true parameters by taking a weighted average of $\mathbf{p_0}$ and $\mathbf{p}$ (see figure 1):

$$\mathbf{p_b} = ({\mathbf{C_0}}^{-1} + \mathbf{C}^{-1})^{-1}({\mathbf{C_0}}^{-1}\mathbf{p_0} + \mathbf{C}^{-1}\mathbf{p}) \tag{2}$$

The estimated covariance matrix of the standard errors for the MAP solution is then:

$$\mathbf{C_b} = ({\mathbf{C_0}}^{-1} + \mathbf{C}^{-1})^{-1} \tag{3}$$

$\mathbf{p_b}$ and $\mathbf{C_b}$ are the parameters that describe the multi-normal distribution $p(a_{\mathbf{p}}|b)$.

2.3 Estimating C.

In order to employ the Bayesian approach, we need to compute $\mathbf{C}$, which is the estimated covariance matrix of the standard errors of the fitted parameters. If the observations are

independent, and each has unit standard deviation, then $\mathbf{C}$ is given by $(\mathbf{A}^T\mathbf{A})^{-1}$. In practice, we don't know the standard deviations of the observations, so we assume that it is equal for all observations, and estimate it from the sum of squared differences:

$$\sigma^2 = \sum_{i=1}^{I} (f(\mathbf{M}\mathbf{x}_i) - p_{13}g(\mathbf{x}_i))^2 \tag{4}$$

This gives a covariance matrix $(\mathbf{A}^T\mathbf{A})^{-1}\sigma^2/(I-J)$, where I refers to the number of sampled locations in the images and J refers to the number of parameters (13 in this case).

However, complications arise because the images are smooth, resulting in the observations not being independent, and a reduction in the effective number of degrees of freedom (from $I-J$). We correct for the number of degrees of freedom using the principles described by Friston (1995a) [although this approach is not strictly correct (Worsley & Friston, 1995), it gives an estimate that is close enough for our purposes]. We can estimate the effective degrees of freedom by assuming that the difference between $\mathbf{f}$ and $\mathbf{g}$ approximates a continuous, zero-mean, homogeneous, smoothed *Gaussian random field*. The approximate parameter of the Gaussian point spread function describing the smoothness in direction d (assuming that the axes of the Gaussian are aligned with the axes of the image coordinate system) can be obtained by (Poline *et al.* , 1995):

$$w_d = \sqrt{\frac{\sigma^2(I-J)}{2\sum_i (\nabla_d(f(\mathbf{M}\mathbf{x}_i) - g(\mathbf{x}_i)))^2}} \tag{5}$$

Typical values for w_d are in the region of 5 to 7 millimeters.

If the images are sampled on a regular grid where the spacing in each direction is s_d, the number of effective degrees of freedom (ν) becomes approximately $(I-J)\prod_d \frac{s_d}{w_d(2\pi)^{1/2}}$, and the covariance matrix can now be estimated by:

$$\mathbf{C} = (\mathbf{A}^T\mathbf{A})^{-1}\sigma^2/\nu \tag{6}$$

Note that this only applies when $s_d < w_d(2\pi)^{1/2}$, otherwise $\nu = I - J$.

2.4 Estimating $\mathbf{p_0}$ and $\mathbf{C_0}$.

The *a priori* distribution of the parameters ($\mathbf{p_0}$ and $\mathbf{C_0}$) was determined from affine transformations estimated from 51 high resolution T1 weighted brain MR images (there were originally 53 images, but two outliers were removed because the registration failed due to poor starting estimates). The subjects were all normal and right handed, and were between 18 and 40 years old (mean age of 25 years). The original group of 53 contained 30 males and 23 females.

The template image used was a high quality T1 image of a single subject that conforms to the space described by Evans *et al.*(1993) (illustrated in figure 2). The basic least squares optimization algorithm was used to estimate these transformations. Each transformation matrix was defined from parameters $\mathbf{q}$ according to:

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & q_1 \\ 0 & 1 & 0 & q_2 \\ 0 & 0 & 1 & q_3 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & cos(q_4) & sin(q_4) & 0 \\ 0 & -sin(q_4) & cos(q_4) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} cos(q_5) & 0 & sin(q_5) & 0 \\ 0 & 0 & & 0 \\ -sin(q_5) & 0 & cos(q_5) & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \cdots$$

$$\cdots \times \begin{pmatrix} cos(q_6) & sin(q_6) & 0 & 0 \\ -sin(q_6) & cos(q_6) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} q_7 & 0 & 0 & 0 \\ 0 & q_8 & 0 & 0 \\ 0 & 0 & q_9 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \times \begin{pmatrix} 1 & q_{10} & q_{11} & 0 \\ 0 & 1 & q_{12} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The results for the translation and rotation parameters (q_1 to q_6) are ignored, since these depend only on the positioning of the subjects in the scanner, and do not reflect variability in head shape and size.

The mean zooms required to fit the individual brains to the space of the template (parameters q_7 to q_9) were 1.10, 1.05 and 1.17 in X (medio-lateral direction), Y (anterior-posterior direction) and Z (dorso-ventral direction) respectively, reflecting the fact that the template was larger than the typical head. Histograms of the values are shown in figure 3. The covariance matrix was:

$$\begin{pmatrix} 0.00210 & 0.00094 & 0.00134 \\ 0.00094 & 0.00307 & 0.00143 \\ 0.00134 & 0.00143 & 0.00242 \end{pmatrix}$$

giving a correlation coefficient matrix of:

$$\begin{pmatrix} 1.00 & 0.37 & 0.59 \\ 0.37 & 1.00 & 0.52 \\ 0.59 & 0.52 & 1.00 \end{pmatrix}.$$

As expected, these parameters are correlated. This allows us to partially predict the optimal zoom in Z given the zooms in X and Y, a fact that is useful for spatially normalizing images containing a limited number of transverse slices.

The means of the parameters defining shear were close to zero (-0.0024, 0.0006 and -0.0107 for q_{10}, q_{11} and q_{12} respectively). The variances of the parameters are 0.000184, 0.000112 and 0.001786, with very little covariance.

Histograms of the values are shown in figure 4. Because of the symmetric nature of the head, there is very little shear in any plane other than the $Y - Z$ plane (parameter q_{12}).

For the Bayesian optimization scheme, the values of $\mathbf{p_0}$ were all set to zero, except for the zoom estimates which were assigned values of 1.10, 1.05 and 1.17. Off-diagonal elements of covariance matrix $\mathbf{C_0}$ were set to zero, with the exception of elements reflecting covariances between zooms. The standard deviations of parameters for translations and rotations were set to arbitrarily high values of 100mm and 30^o.

2.5 Incorporating the Bayesian Approach into the Optimization.

As mentioned previously, when the parameter estimates are close to the minimum the registration problem is almost linear. Prior to this, the problem is non-linear and covariance matrix $\mathbf{C}$ no longer directly reflects the certainties of the parameter estimates. However, it does indicate the certainties of the changes made in the parameter estimates at each iteration, so this information can still be incorporated into the iterative optimization scheme.

By combining Eqns. (1), (2) and (6), we obtain the following scheme:

$$\mathbf{p_b}^{(n)} = (\mathbf{C_0}^{-1} + \alpha)^{-1}(\mathbf{C_0}^{-1}\mathbf{p_0} + \alpha\mathbf{p_b}^{(n-1)} - \beta) \tag{7}$$

where $\alpha = \mathbf{A}^T\mathbf{A}\nu/\sigma^2$ and $\beta = \mathbf{A}^T\mathbf{b}\nu/\sigma^2$.

Another way of thinking about this optimization scheme, is that two criteria are simultaneously being minimized. The first is the sum of squares difference between the images, and the second is a scaled distance between the parameters and their known expectation.

2.5.1 Stopping Criterion.

The optimal solution is no longer that which minimizes the sum of squares of the residuals, so the rate of change of σ^2 is not the best indication of when the optimization has converged. The objective of the optimization is to obtain a fit with the smallest errors. These errors are described by the covariance matrix of the parameter estimates, which in the case of this optimization scheme is $(\alpha + {\mathbf{C_0}}^{-1})^{-1}$. The 'tightness' of the fit is reflected in the determinant of this matrix, so the optimal solution should be achieved when the determinant is minimized. In practice we look at the rate of change of the log of the determinant.

3 Results and Discussion.

3.1 Plots of convergence - with and without Bayesian extension.

The algorithm was applied to 100 T1 weighted images, in order to match the images to a T1 template image. All images were smoothed with a Gaussian kernel of 8mm full width at half maximum. The voxels were reduced to $2 \times 2 \times 4$mm with a field of view of $256 \times 256 \times 128$mm in X, Y and Z respectively, in order to facilitate faster computations.

The optimizations were performed three times: (A) Without the Bayesian scheme, for a 12 parameter affine transformation. (B) With the Bayesian scheme, for a 12 parameter affine transformation. (C) Without the Bayesian scheme, for a six parameter rigid body transformation (to demonstrate that the Bayesian scheme is not simply optimizing a rigid body transformation).

During the optimization procedure, the images were sampled approximately every 8mm. 32

iterations were used, and the value of σ^2 recorded for each iteration. Although we do not propose that convergence should be indicated by σ^2, it provides a useful index to demonstrate the relative performance. 50 of the subjects were given good starting estimates (i), and 50 were given starting estimates that deviated from the optimal solution by about 10cm (ii).

There were 2 cases from (ii) in which the starting estimates were insufficiently close to the solution, for either (A) or (B) to converge satisfactorily. These cases have been excluded from the results.

Figure 5 shows the average σ^2 for all images plotted against iteration number. As can be seen from these plots, (B) leads to a more rapid estimation of the optimal parameters, even though convergence appears faster at the start of (A). The plot of convergence for (C) illustrates the point that the Bayesian method is not over-constrained and simply optimizing a set of rigid body parameters.

Figure 6 compares the number of iterations required by (A) and (B) in order to reduce the σ^2 to within 1% of the minimum of both schemes. In several cases of (A), the optimization had not converged within the 32 iterations. There were only 5 cases where (B) does not obtain a value for σ^2 that is as low as that from (A). In two of the cases, the results from (A) were very close to those from (B). However, in the other three cases, examination of the parameter estimates from scheme (A) showed that it had found a minimum that was clearly not a proper solution. The zooms determined, after 32 iterations, were (0.96,0.98,0.11), (2.10,0.72,0.0003) and (1.09,0.24,0.02). These are clearly not correct!

The algorithm requires relatively few iterations to reach convergence. The speed of each iteration depends upon the number of sampled voxels. On a SPARC Ultra 2, an iteration takes one second when about 26000 points are sampled.

3.2 Comparisons of affine normalization with limited data.

Occasionally the image that is to be spatially normalized is of poor quality. It may have a low signal to noise ratio, or it may contain only a limited number of slices. When this is the case, the parameter estimates for the spatial normalization are likely to be unreliable. Here we present a further comparison of affine registrations with and without the incorporation of prior information [(E) and (D) respectively]. This time, we sampled only four planes from the images, to simulate an effective field of view of 16 mm. The optimizations were given good initial parameter estimates, and the results compared with those obtained using the complete data.

The resulting parameter estimates from (D) and (E) are plotted against those from (B) in figure 7. As can be seen from the plots, where the parameters can be estimated accurately, the results from (D) and (E) are similar. However, where there is not enough information in the images to determine an accurate parameter estimate, the results of (E) are properly biased towards the prior estimate.

3.3 Implications for Nonlinear Warping.

A Bayesian approach to non-linear image registration is nothing new. The incorporation of prior knowledge about the properties of the allowed warps is fundamental to all successful non-linear registration approaches. Gee *et al.*(1995) have already described one Bayesian approach to non-linear image registration.

For the non-linear spatial normalization of brain images prior to statistical analysis, the objective is to warp the images such that homologous regions of different brains are moved as close together as possible. A high number of parameters are required to encompass the range of possible non-linear warps. With many parameters relative to the number of independent observations, the errors associated with the fit are likely to be very large. The use of constraints (such as preserving a one to one mapping between image and template) can reduce these errors, but they still remain considerable. For this purpose, the simple

minimization of differences between the images is not sufficient. Although the normalized images may appear similar to each other, the data may in-fact have been 'over-fitted', resulting in truly homologous regions being moved further apart. Other researchers circumvent this over-fitting problem by restricting their spatial normalization to just an affine transformation. A Bayesian approach similar to that described here would attempt to reach an optimum compromise between these two extremes.

Although the incorporation of an optimally applied MAP approach into non-linear registration should have the effect of biasing the resulting deformations to be smoother than the true deformations, it is envisaged that homologous voxels would be registered more closely than for unconstrained deformations. The measurements above demonstrate that brain lengths vary with a standard deviation of about 5% of the mean. A suitable starting point may be to assume that there is roughly the same variability in the lengths of the different brain substructures. The relative sizes of voxels before and after spatial normalization is reflected in the derivatives of the fields that describe the deformation. Therefore, an improved non-linear spatial normalization may be achieved by assigning a prior that these derivatives should have a standard deviation of about 0.05. For deformations that are defined by a linear combination of smooth basis functions (Friston *et al.* , 1995b), the derivatives of the deformations are simply the same linear combination of the derivatives of the basis functions. It is therefore possible to assign a covariance matrix describing a prior distribution for the coefficients of the transformation.

An alternative approach would be to assume that the relative voxel volumes are drawn from a known log-normal distribution (and therefore are always positive). These relative volumes are described by the determinants of the Jacobian of the deformation field. An assumption of this type will ensure that there will always be a one-to-one mapping in the deformation, and so should be more robust for estimating higher resolution non-linear deformations.

The above two models assume that every voxel has similar deformable properties (see Thompson *et al.*(1996) to assess the validity of this assumption). However, they can both be ex-

tended by incorporating information on the distribution of deformation fields from a number of different subjects. The first of the two models could incorporate a covariance matrix computed from real data, whereas the second could utilize some representation of the variability of voxel sizes, in the form of an image (or series of images).

The simple affine transformation was chosen for this project, as the variability in brain dimensions is simple to characterize as a multi-normal distribution. Unfortunately, the full characterization of a probability density function describing the *a priori* distribution of non-linear warps is not so straightforward. Thompson *et al.*(1996) have aleady begun to characterize normal morphological variability of the brain, in order to identify structural abnormalities. The variability was derived by estimating non-linear registrations for a number of images using a fluid model, and is represented by the means and variances of the displacements at each voxel. This representation is able to encode some of the parameters describing normal variability, but much of the information is inevitably lost. Le Briquer and Gee (1997) use a global model to represent the normal variability of the deformations. The model is constructed from the principal components of a number of previously estimated deformation fields, and allows the incorporation of spatial correlations in the warps. The next challenge will be to determine an optimum compact form to describe the structural variability, and also to estimate the parameters describing the distribution.

References

Briquer, L. Le, & Gee, J. C. 1997. Design of a statistical model of brain shape. *Pages 477–482 of: Information processing in medical imaging.*

Christensen, G. E., Rabbitt, R. D., & Miller, M. I. 1996. Deformable templates using large deformation kinematics. *IEEE transactions on image processing*, **5**, 1435–1447.

Evans, A. C., Collins, D. L., Mills, S. R., Brown, E. D., Kelly, R. L., & Peters, T. M. 1993. 3D statistical neuroanatomical models from 305 MRI volumes. *Pages 1813–1817 of: Proc. IEEE-nuclear science symposium and medical imaging conference.*

Friston, K. J., Holmes, A. P., Poline, J.-B., Grasby, P. J., Williams, S. C. R., Frackowiak, R. S. J., & Turner, R. 1995a. Analysis of fMRI time series revisited. *NeuroImage*, **2**, 45–53.

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., & Frackowiak, R. S. J. 1995b. Spatial registration and normalization of images. *Human brain mapping*, **2**, 165–189.

Gee, J. C., Briquer, L. Le, & Barillot, C. 1995. Probablistic matching of brain images. *Pages 113–125 of: Information processing in medical imaging.*

Poline, J.-B., Friston, K. J., Worsley, K. J., & Frackowiak, R. S. J. 1995. Estimating smoothness in statistical parametric maps: Confidence intervals on p-values. *J. comput. assist. tomogr.*, **19(5)**, 788–796.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992. *Numerical recipes in C (second edition).* Cambridge: Cambridge.

Talairach, J., & Tournoux. 1988. *Coplanar stereotaxic atlas of the human brain.* New York: Thieme Medical.

Thompson, P. M., & Toga, A. W. 1996. Visualization and mapping of anatomic abnormalities using a probablistic brain atlas based on random fluid transformations. *Pages 383–392 of: Proceedings of the international conference on visualization in biomedical computing.*

Thompson, P. M., Schwartz, C., Lin, R. T., Khan, A. A., & Toga, A. W. 1996. 3D statistical analysis of sulcal variability in the human brain. *Journal of neuroscience*, **16(13)**, 4261–4274.

Worsley, K. J., & Friston, K. J. 1995. Analysis of fMRI time-series revisited - again. *NeuroImage*, **2**, 173–181.

Acknowledgments

This work was supported by the Wellcome Trust and carried out while at the Montreal Neurological Institute. All MRI data was from the Montreal Neurological Institute.

Figure Legends

Figure 1

This figure illustrates a hypothetical example with one parameter. The solid Gaussian curve (a) represents the *a priori* probability distribution (p.d.f), and the dashed curve (b) represents a parameter estimate (from fitting to observed data) with its associated certainty. We know that the true parameter was drawn from distribution (a), but we can also estimate it with the certainty described by distribution (b). Without the MAP scheme, we would probably obtain a more precise estimate for the true parameter by taking the most likely *a priori* value, rather than the value obtained from a fit to the data.

The dotted line (c) shows the p.d.f that would be obtained from a MAP estimate. It combines previously known information with that from the data to give a more precise estimate.

Figure 2

An illustration of the template (left) and an affine registered image (right). Note that the template is pre-smoothed to facilitate faster matching.

Figure 3

The distribution of the zooms in X, Y and Z required to fit 51 brains to a standard space. The figure shows a histogram of the values, and also the Gaussian curve that best fits the distribution.

Figure 4

The distribution of the shears required to fit the 51 brains to a standard space.

Figure 5

The average σ^2 for the images plotted against iteration number. Left: given good starting estimates (i). Right: given poor starting estimates (ii). The dashed lines (A) show convergence for a 12 parameter affine transformation without using the Bayesian scheme. The solid lines (B) show the same, but with the Bayesian scheme. Convergence for a six parameter rigid body transformation (C) is shown in the dotted lines.

Figure 6

The number of iterations in which convergence to within 1% of the minimum mean residual sum of squares had not been reached. The non-Bayesian scheme (A) is on the X axis, with the Bayesian scheme (B) on the Y axis. Results from optimizations given good starting estimates are shown as circles, whereas those with bad starting estimates are shown as crosses.

Figure 7

Plots of the parameter estimates from reduced data, against estimates using the complete data. As expected, the Bayesian scheme makes little difference for the estimates of the zoom in the X direction [(a) and (b)], whereas the Bayesian scheme heavily biases the zoom in Z towards the mean of the prior distribution [(c) and (d)].

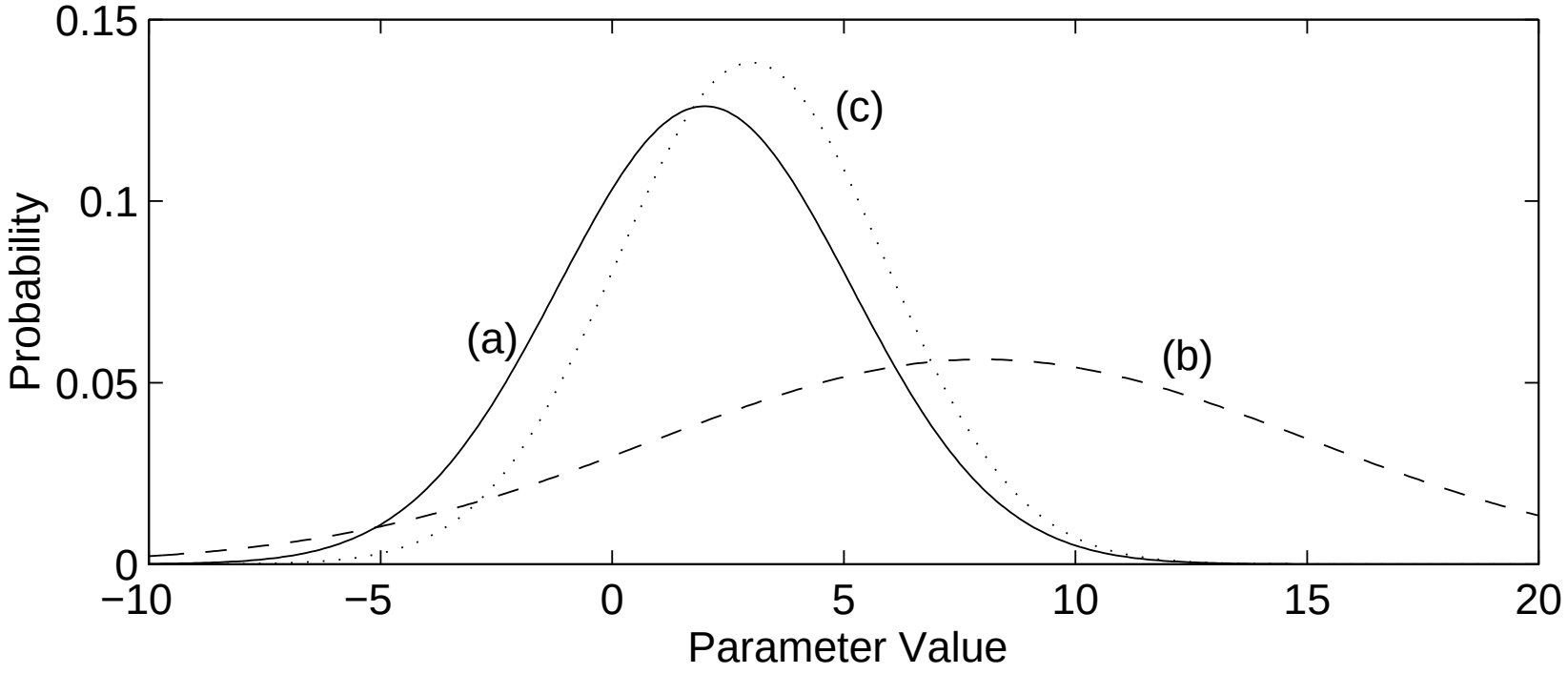

Figure 1:

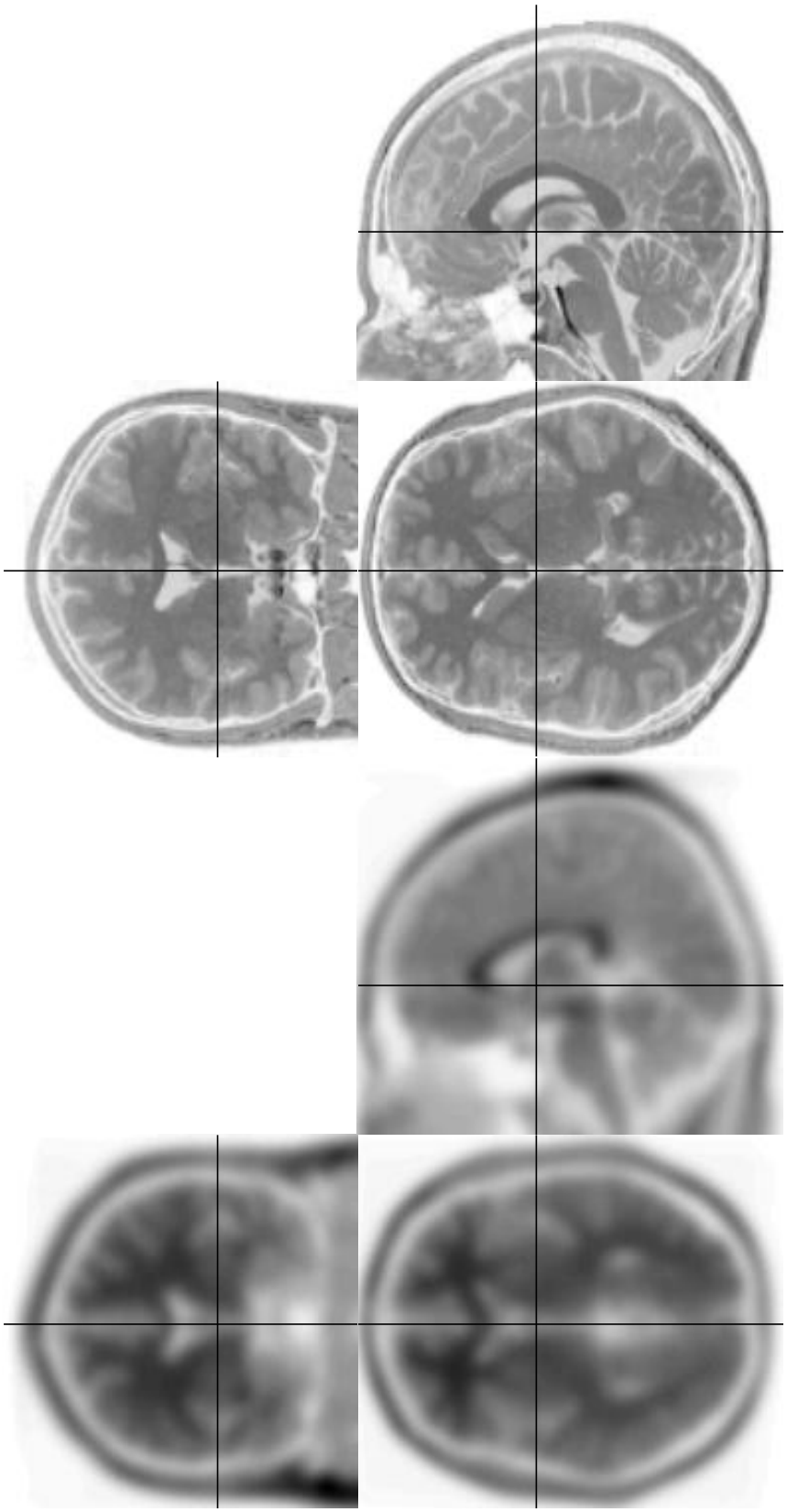

Figure 2:

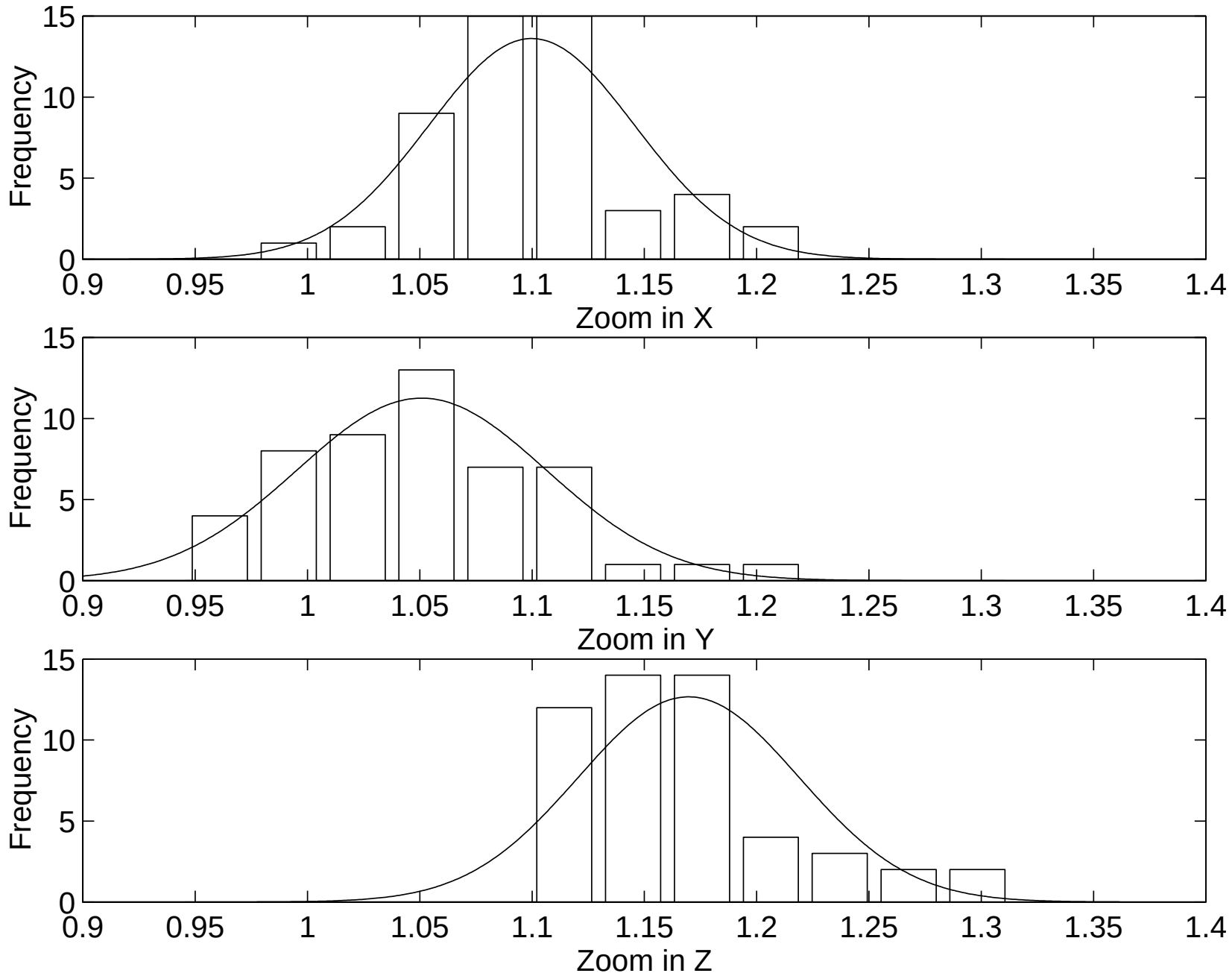
Frequency
15
10
5
0
0.9
0.95
1
1.05
1.1
1.15
1.2
1.25
1.3
1.35
1.4
Zoom in X
Frequency
15
10
5
0
0.9
0.95
1
1.05
1.1
1.15
1.2
1.25
1.3
1.35
1.4
Zoom in Y
Frequency
15
10
5
0
0.9
0.95
1
1.05
1.1
1.15
1.2
1.25
1.3
1.35
1.4
Zoom in Z

Figure 3:

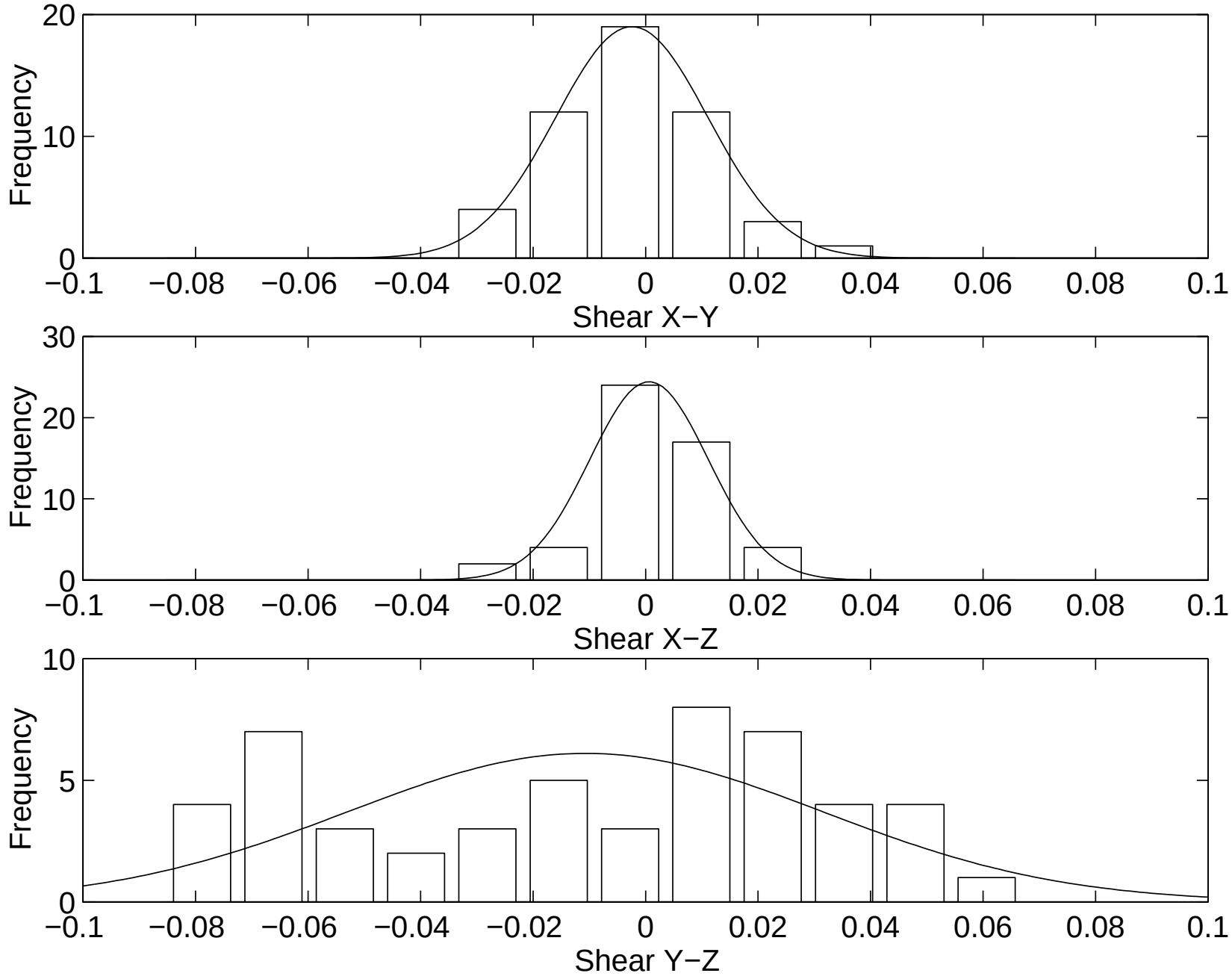
Frequency
20
10
0
−0.1
−0.08
−0.06
−0.04
−0.02
0
0.02
0.04
0.06
0.08
0.1
Shear X−Y
Frequency
30
20
10
0
Shear X−Z
Frequency
10
5
0
Shear Y−Z

Figure 4:

Convergence
Mean Residual Squares
Iteration #
(A) − dashed
(B) − solid
(C) − dotted
2000
1500
1000
500
0
10
20
30
Convergence
Mean Residual Squares
Iteration #
(A) − dashed
(B) − solid
(C) − dotted
2000
1500
1000
500
0
10
20
30

Figure 5:

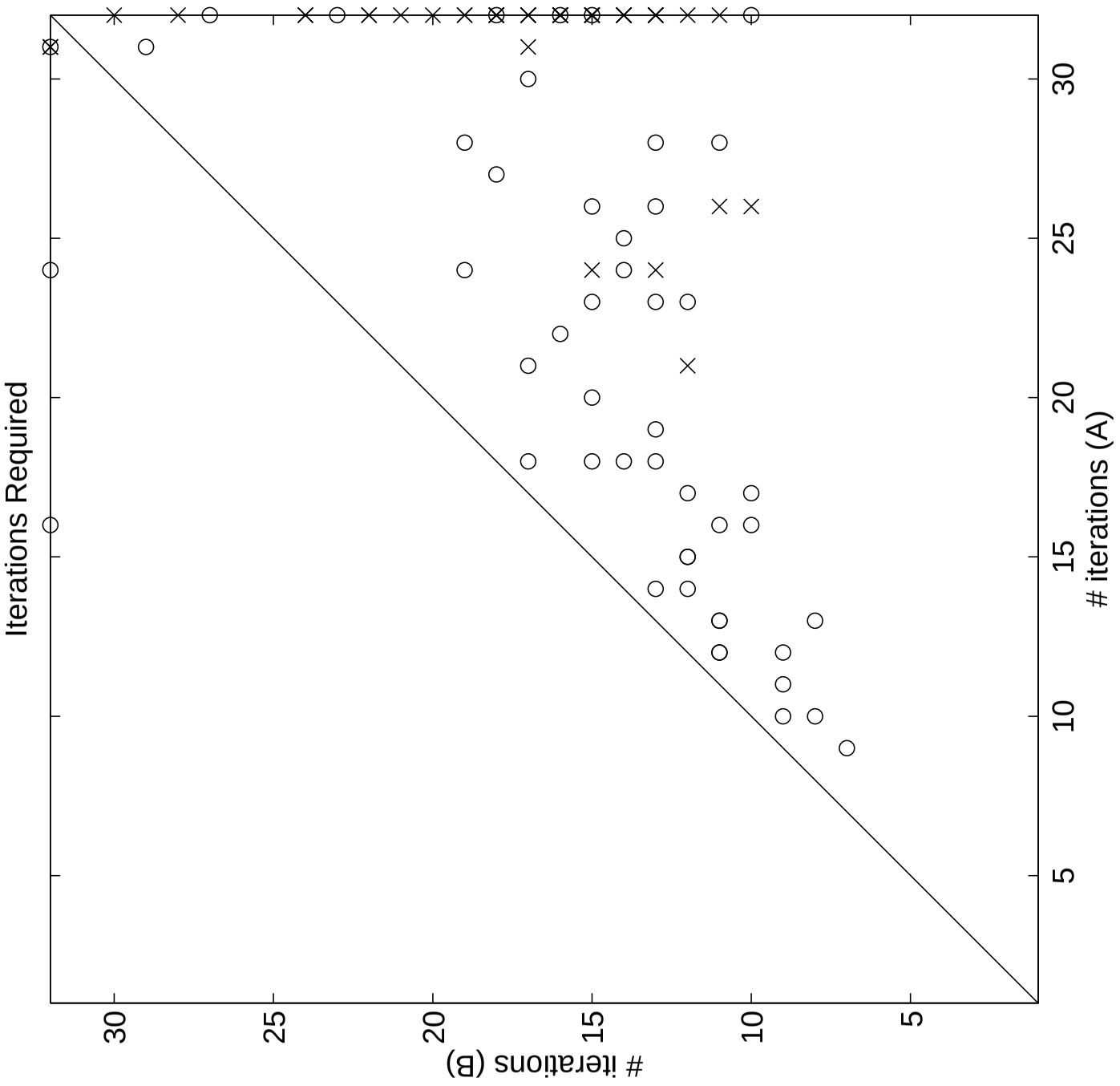
Iterations Required
iterations (B)
iterations (A)
5
10
15
20
25
30

Figure 6:

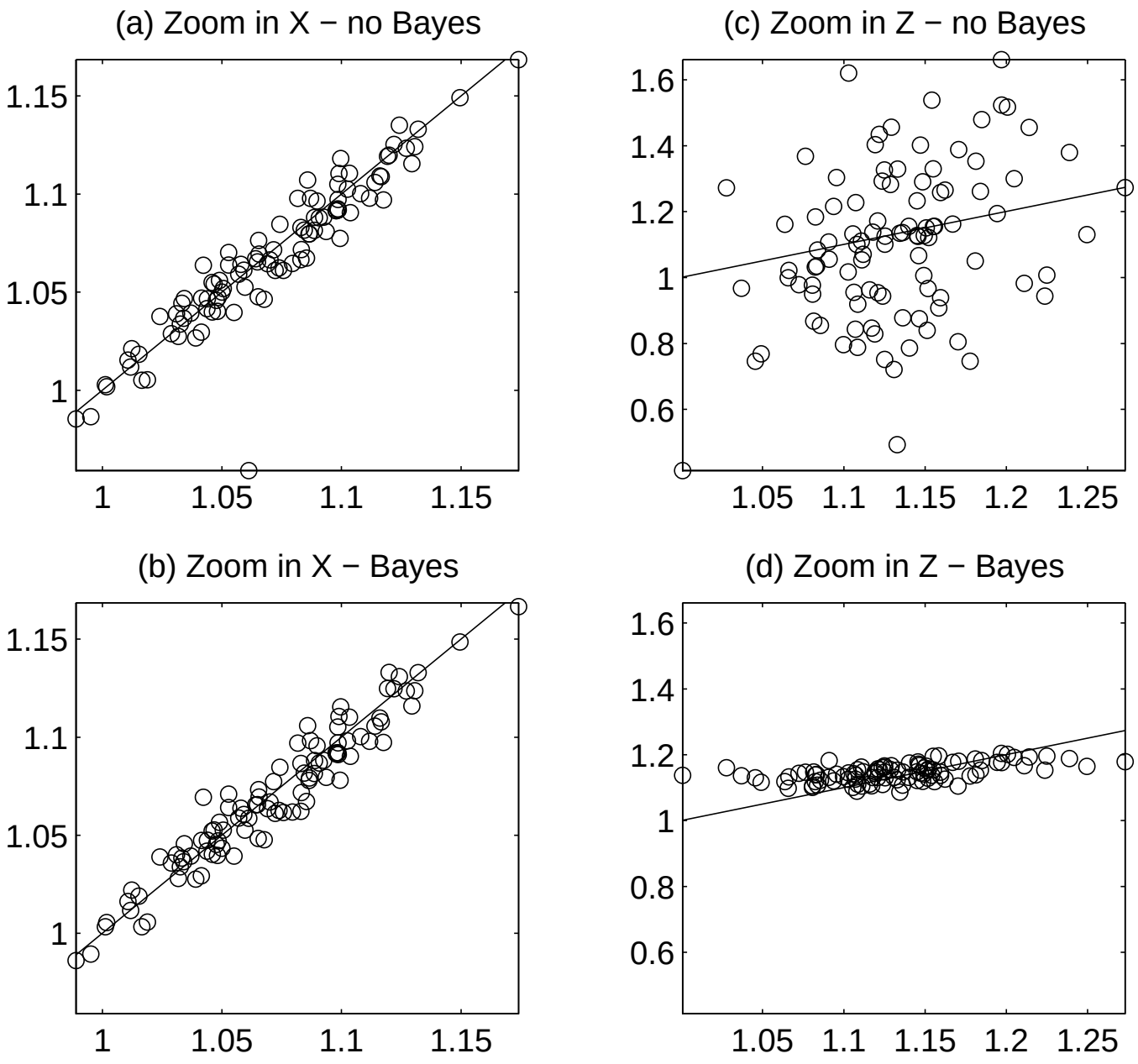
(a) Zoom in X − no Bayes
(c) Zoom in Z − no Bayes
(b) Zoom in X − Bayes
(d) Zoom in Z − Bayes
1.15
1.1
1.05
1
1.6
1.4
1.2
0.8
0.6
1.05
1.1
1.15
1.2
1.25

Figure 7: