Noise in bilinear problems

J.A.Haddon D.A. Forsyth

Computer Science Division
U.C. Berkeley
Berkeley, CA 94720
{haddon,daf}@cs.berkeley.edu

Abstract: *Despite the wide application of bilinear problems to problems both in computer vision and in other fields, their behaviour under the effects of noise is still poorly understood. In this paper, we show analytically that marginal distributions on the solution components of a bilinear problem can be bimodal, even with Gaussian measurement error. We demonstrate and compare three different methods of estimating the covariance of a solution. We show that the Hessian at the mode substantially underestimates covariance.*

Many problems in computer vision can be posed as bilinear problems: i.e. one must find a solution to a set of equations of the form

$$c_k = \sum_{ij} g_{ijk} a_i b_j$$

for c_k a set of known terms (henceforth **measurements**), and g_{ijk} a set of known interaction terms. Typically, a_i and b_j are constrained in some way to allow a unique solution. The most familiar example is Tomasi and Kanade's formulation of orthographic structure-from-motion [9]; shape-from-shading and other vision problems can be framed this way too (see [5] for a review. Other naturally bilinear problems include: inverse kinematics for parallel manipulators [6]; and molecular conformation [1]).

The effect of noise in the measurements is not well understood. Figure 1 shows a scatter plot of of point positions reconstructed from an orthographic image sequence with Gaussian noise. Not only are the distributions quite obviously not Gaussian, some even appear bimodal. *There is no reason to expect that they should be Gaussian*. As we shall see, noise can lead to bimodal marginal posteriors on a_i, meaning that straightforward covariance estimates are extremely unreliable.

This paper compares three methods of estimating covariance for marginals on a_i and b_j in bilinear problems. In section 1, we analyse some simple examples which illustrate the problem. We then examine three possible ways of estimating a covariance, and show that two which appear in the literature can be rather misleading in their estimates of covariance. We focus on the orthographic structure-from-motion problem because it is most familiar, but the conclusions that we draw are equally applicable to any bilinear problem.

1 Analytical examples

Even quite simple examples display considerable complexity, but have the advantage that analysis is possible.

1.1 A One-Dimensional Example

Consider a 2×2 measurement matrix $\mathcal{D}$, which is assumed to be close to a rank-1 matrix, differing only by a Gaussian noise matrix $\mathcal{W}$, with $w_{ij} \sim N(0, \sigma^2)$. We can write this as:

$$\begin{bmatrix} 1 \\ u \end{bmatrix} \begin{bmatrix} x & y \end{bmatrix} = \mathcal{D} + \mathcal{W}$$

We have constrained the first component of the column vector to be 1 to remove the scaling ambiguity that would otherwise be present.

Since the noise is iid Gaussian, we can easily write the posterior pdf:

$$P(u, x, y|\mathcal{D}) \propto \exp\left[-\frac{(x-d_{11})^2+(y-d_{12})^2+(ux-d_{21})^2+(uy-d_{22})^2}{2\sigma^2}\right]$$

Note that this distribution is *not* jointly Gaussian on u, x and y. However, given some constant u, the conditional *is* jointly Gaussian on x and y. We can use this fact to calculate the marginal distribution on u by integrating out x and y.

$$P(u|\mathcal{D}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(u, x, y|\mathcal{D}) dx dy$$

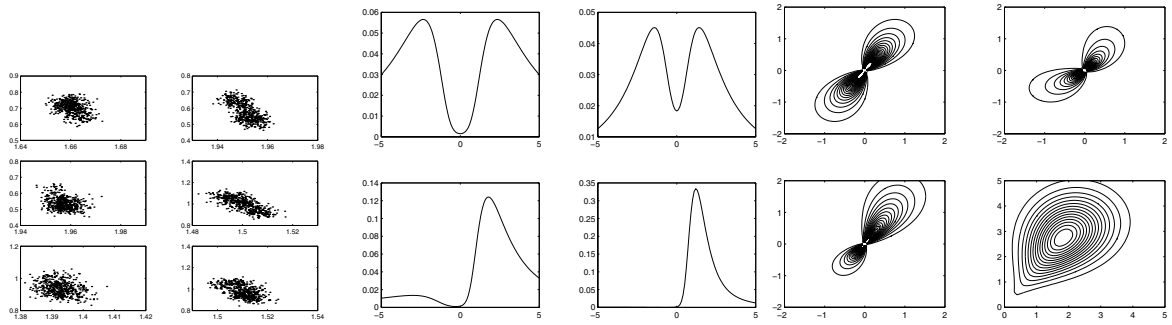

Figure 1: *One should not assume variables are distributed according to a Gaussian without showing that it is justified—there are many distributions which cannot be modelled by a Gaussian. Pictured* **left** *are samples of reconstructed point positions, projected onto a plane parallel to the optical axis. None of these figures show a Gaussian distribution; the figures on the right are even bimodal. Estimating the covariance of these scatter plots using the Hessian at the mode is going to lead to serious problems, because large chunks of probability will not be accounted for. Marginal distributions in bilinear problems can be bimodal, even in very simple problems. Pictured* **center** *are the marginal distributions on u for the problem of section 1.1, assuming Gaussian measurement error. Even under this assumption, in three of the four cases, we have two maxima in the marginal probability density function. Even the marginal distributions on point positions have strange shapes. Pictured* **right** *are the marginals on x, y for the same four data matrices as in figure 1. While these distributions are not bimodal, a Gaussian approximation will necessarily miss large regions of probability.*

$$\propto \frac{\exp\left[-\frac{(d_{11}^2+d_{12}^2)u^2-2(d_{11}d_{21}+d_{12}d_{22})u+(d_{21}^2+d_{22}^2)}{2\sigma^2(1+u^2)}\right]}{1+u^2}$$

To find the critical points of this function, we differentiate with respect to u to obtain the product of a rational function and an exponential. The numerator of the rational function is a cubic, which means that there may be up to three critical points. Figure 1 shows the marginal distributions on u for four different data matrices. In three of these cases, the marginals are actually bimodal; in the fourth, the marginal is unimodal, but very strongly asymmetric.

In an analogous fashion, we can find the marginal on x and y; the contour plots in figure 1 depict these distributions for the same data matrices as in figure 1. These are very strange distributions, although none of them are technically bimodal, since the maximum value in the first three cases is at the origin. (The value at the origin is actually undefined, but the probability increases as we approach the origin.) Nevertheless, these distributions will not be well-described by a Gaussian (the contours for a Gaussian are concentric ellipses).

Note that the property of being bimodal is *not* dependent on the parametrisation for a diffeomorphic change of parametrisation. Fixing the first component of the column vector at 1 is one choice of parametrisation, but we could have chosen any equivalent parametrisation—the stationary points of the distribution will remain stationary. This means that the bimodality we observe here is not merely an artefact of the parametrisation, but is inherent to the problem. In particular, if we chose some affine transformation of the factors (as is often done in order to satisfy certain constraints, for example, requiring the camera to be orthographic) we will observe the same qualitative behaviour.

This relatively simple example has shown that bilinear problems are able to produce alarming behaviour. Even a rank-1 matrix corrupted by Gaussian noise gives us bimodal marginals. Bimodal marginals are problematic for several reasons: first, they cannot be approximated as easily (*e.g.* by a single Gaussian). Second, if we are not aware of the multimodality, and simply maximise the probability, we may miss an entire region of space with significant density, even if we do successfully find the large mode. If we recognise the possibility of multiple modes, we can make our system robust to their effects.

1.2 A Two-Dimensional Example

Now let us consider a two-dimensional case. Let us suppose that we have a 3×3 rank-2 data matrix, corrupted by Gaussian noise. We can write:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ u_1 & u_2 \end{bmatrix} \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{bmatrix} = \mathcal{D} + \mathcal{W}$$

where $\mathcal{W}$ is a Gaussian noise matrix again. The noise is iid Gaussian, and the conditional on u_1 and u_2 is jointly Gaussian in the x_i and y_i. The marginal distribution is thus easily calculated.

Figure 2 shows a plot of $P(u_1|D, u_2 = 0)$ and surface plot of $P(u_1, u_2|D)$. It is clear from the figure that the marginal distribution is not Gaussian. We shall see in section 2.1 that if we simply estimate the distribution by ex-

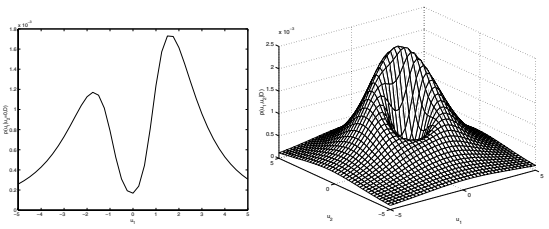

Figure 2: *Marginal distributions in bilinear problems are not in general Gaussian. The problem depicted here is discussed in section 1.2.* **Left:** *The conditional distribution $p(u_1|u_2 = 0, \mathcal{D})$. Note the presence of two minima.* **Right:** *The marginal $p(u_1, u_2|\mathcal{D})$. There is a unique maximum, but there is also a saddle and a local minimum. We have assumed measurements with Gaussian noise, and even so, we see that the distribution is very complex. There is a region of low probability at the origin, surrounded by regions of higher probability. This is likely to cause significant underestimation of the covariance of the distribution. Note that this is the marginal on two dimensions of a general six-dimensional space—we have fixed the other four dimensions so that we do not need to consider the ambiguity.*

amining its behaviour near the maximum, this will cause a covariance estimate which is much smaller than reality.

2 Methods of estimating a covariance

Given a set of image measurements, we would like to find a good solution for the camera parameters and point positions which could have generated these images. But such a solution is of little value if it has a huge variance—if the solution we find is only slightly better than any randomly chosen solution, then the claim that this is the "best" solution is a rather vacuous one. Therefore, we would like to be able to describe the covariance of the camera parameters and point positions, given the measurements. We have now seen several reasons why the covariance of the solution to a bilinear problem is difficult to estimate. We now turn our attention to the problem of actually estimating the covariance.

2.1 Laplace's approximation

One common approach for dealing with complex covariances is to assume that they are Gaussian, an assumption which, as we have seen, is often unjustified. This is known in the numerical integration literature as Laplace's approximation [2]. When integrating a function, we care mostly about where the function is big; the details of its shape are less important. It is this observation that allows us to approximate a unimodal distribution as a Gaussian. However, if the distribution is bimodal, or if it is hard to tell how large the mode is, a Gaussian approximation can be very poor.

To find the Gaussian approximation, we note that, for a true Gaussian distribution, the mean and mode are equal. We therefore set the mean of the Gaussian approximation to the mode of the distribution. To find the covariance, we note that, if $g(\mathbf{x})$ is a Gaussian distribution,

$$\begin{aligned} g(\mathbf{x}) &\propto \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right] \\ \log g(\mathbf{x}) &= k - \frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) \\ H &= -\frac{\partial^2 \log g(\mathbf{x})}{\partial \mathbf{x}\partial \mathbf{x}^T} = \Sigma^{-1} \end{aligned}$$

For a true Gaussian, the inverse covariance matrix is equal to the Hessian of the negative logarithm. For our approximation, we evaluate the Hessian of the negative log of the distribution at the mode, and take this to be the inverse covariance matrix.

If we wish to examine the marginals on just a few variables, it is important to note that the covariance of the marginal distribution is obtained by extracting the relevant rows and columns from the full covariance matrix, and not by inverting only the relevant block of the inverse covariance matrix. If the covariance matrix has a block diagonal structure, then, of course, these two are equivalent, but any bilinear problem will have a strong off-diagonal component, due to the presence of terms with the product of two variables (see figure 3).

If we pretend that the us and vs are independent (a fallacy), then the inverse covariance on the vs is block diagonal, and therefore easy to invert. Note that this will yield a smaller covariance than that obtained by inverting the full Hessian, because it corresponds to fixing the cameras, and then estimating the point positions; if we also allow the cameras to move, we expect the point positions will vary more.

One problem with Laplace's approximation is that the Hessian can seriously overestimate the covariance of a distribution if that distribution is rather flat at its peak. Consider the distribution $f(x) = k\exp[-\frac{1}{2}x^4]$ depicted by the solid line in figure 3. This distribution clearly has a mean of zero, and we can integrate numerically to find that the variance is 0.478. (The dotted line depicts a Gaussian distribution with zero mean and this variance—not an unreasonable approximation.) Laplace's approximation will also put the mean at zero (this is the mode of the distribution), but when we evaluate the Hessian at the mode in order to find the covariance, we run into problems; the variance is infinite, and all points are equally likely. Note that the structure-from-motion problem does in fact have fourth-order terms in the Hessian, because the exponent contains the square of the product of camera parameters and point positions. This suggests that the structure-from-

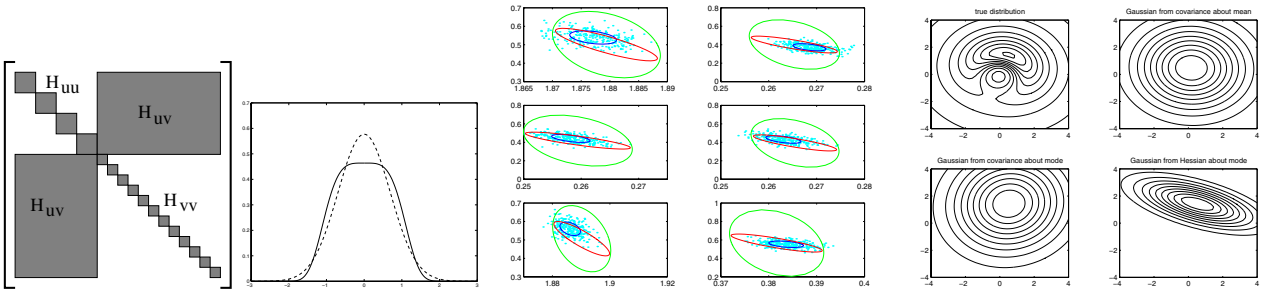

Figure 3: **Left:** *Schematic structure of the Hessian. White regions indicate zero entries in the matrix, while the grey regions represent entries which are in general nonzero. The camera parameters (resp. points) at one frame are independent of camera parameters (resp. points) at other frames. Off diagonal terms indicate that the camera parameters (resp. points) do depend on the positions of all of the points (resp. camera parameters).* **Center left:** *Laplace's approximation can give extremely bad estimates for very flat distributions. The solid line depicts the density* $k\exp[-\frac{1}{2}x^4]$*, while the dotted line depicts a Gaussian with the same mean and variance. The Laplace approximation is a Gaussian with zero mean and infinite variance, and is therefore not pictured.* **Center right:** *Laplace's approximation severely overestimates the variance of the distribution (in each case, it yields the outside ellipse). If we assume independence of point positions and camera parameters, we obtain a better estimate from Laplace's approximation (in each case, the second smallest ellipse). The dots are samples of marginal point positions for the hotel sequence, and the interior ellipse indicates the covariance of these samples.* **Right:** *A marginal distribution from a toy bilinear problem (see section 1.2) and three Gaussian approximations. (upper right) Estimating the covariance about the mean as* $E((\mathbf{x}-\mu)(\mathbf{x}-\mu)^T)$*. (lower left) Estimating the covariance about the mode as* $E((\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T)$*. (lower right) Laplace approximation, estimating the covariance from the Hessian at the mode. The Laplace approximation greatly underestimates the covariance.*

motion problem may be susceptible to this difficulty.

In figure 3, we show the marginal distributions on six points from the hotel sequence.[1] Note that the Laplace approximation significantly overestimates the variance, for the reason we have already noted. The independence assumption decreases this variance, but is still an overestimate.

2.1.1 Zero eigenvalues

Another issue which arises when attempting to find Laplace's approximation for a bilinear problem is that the Hessian will have several zero eigenvalues, due to the presence of ambiguities (also known as gauge freedoms). In the structure-from-motion problem, there is an arbitrary scale and an arbitrary choice of frame (rotation and translation). There is thus a seven dimensional manifold corresponding to the actions of these arbitrary choices of scale and frame that preserve the value of the posterior. This means that the Hessian will have seven zero eigenvectors. By fixing the scale and frame (*i.e.* removing seven rows and columns from the Hessian) we can eliminate the singular directions and invert the matrix. If our parametrisation of the problem were gauge independent (see [8]), removing these rows and columns would be unnecessary, but it is necessary to account for these ambiguities in some fashion. If we do not take this effect into account, we will be attempting to invert a singular matrix in order to find the covariance matrix. For a real problem, this matrix may not be exactly singular, but there will still be numerical issues that arise when we invert it.

[1] The hotel sequence is courtesy of the Modeling by Videotaping group in the Robotics Institute, Carnegie Mellon University.

2.2 Morris and Kanade's method

Another method to estimate a covariance is to examine the maximum likelihood estimates corresponding to samples of perturbed data. In [7], Morris and Kanade take the measurements, perturb them by Gaussian noise, and find the maximum likelihood estimate of the camera parameters and point positions for the perturbed measurement. They then repeat this many times to obtain many samples and then use this as their "ground truth". Because the Morris-Kanade method uses samples, it does have the expressive power to represent an arbitrary distribution (given enough samples) unlike the Laplace approximation, which assumes *a priori* that the distribution is Gaussian. Unfortunately, the procedure is only valid in the special case of linear transformations.

Let us consider a problem in which we are attempting to estimate the state x, based on our observation of the random variable y given by $y = \phi(x) + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$. We shall assume that ϕ is invertible. Then the distribution

of x given y is

$$p(x|y) = k \exp\left[-\frac{1}{2\sigma^2}(\phi(x) - y)^2\right]$$

To find k, we note that $\int p(x|y)dx = 1$, so that

$$\begin{aligned} k &= \left(\int \exp\left[-\frac{1}{2\sigma^2}(\phi(x) - y)^2\right] dx\right)^{-1} \\ &= \left(\int \frac{e^{-\frac{z^2}{2\sigma^2}}}{\phi'(\phi^{-1}(z+y))} dz\right)^{-1} \end{aligned}$$

where we have made the substitution $z = \phi(x) - y$. Using the same substitution, we can write expressions for the true conditional expectations of x and x^2:

$$\begin{aligned} E[x|y] &= k \int \frac{\phi^{-1}(y+z)e^{-\frac{z^2}{2\sigma^2}}}{\phi'(\phi^{-1}(y+z))} dz \\ E[x^2|y] &= k \int \frac{[\phi^{-1}(y+z)]^2 e^{-\frac{z^2}{2\sigma^2}}}{\phi'(\phi^{-1}(y+z))} dz \end{aligned}$$

Morris and Kanade's method takes the measured value of y, and adds Gaussian noise to it to obtain samples of y_i. Given y_i, we find the maximum likelihood estimate x_i, and consider these to be samples from the distribution $p(x|y)$. Let us examine this claim. We write $y_i = y + \zeta_i$ where $\zeta_i \sim N(0, \eta^2)$ and $x_i = \phi^{-1}(y_i) = \phi^{-1}(y + \zeta_i)$ Note that we need not necessarily take $\eta = \sigma$; in fact, often we will not know the exact value of σ.

Calculating expectations,

$$\begin{aligned} E[x_i|y] &= \frac{1}{\eta\sqrt{2\pi}} \int \phi^{-1}(y+\zeta) e^{-\frac{\zeta^2}{2\eta^2}} d\zeta \\ E[{x_i}^2|y] &= \frac{1}{\eta\sqrt{2\pi}} \int [\phi^{-1}(y+\zeta)]^2 e^{-\frac{\zeta^2}{2\eta^2}} d\zeta \end{aligned}$$

Comparing the expressions for $E[x|y]$ and $E[x_i|y]$, we see that the integrand differs by a factor of $\phi'(\phi^{-1}(z+y))$ in the denominator. Thus, the Morris-Kanade method does not give an unbiased estimate of the mean (or the variance) of a general distribution. If we take $\phi(x) = x^2$, $y = 1$, $\sigma = \eta = 0.2$, this method overestimates the mean by 1.1%, and underestimates the variance by 5%. If we choose η so that $E[y_i|x]$ is an unbiased estimator of the mean ($\eta = 0.336$), we will overestimate the variance by a factor of 3!

If ϕ is linear, then the factor $\phi'(\phi^{-1}(z+y))$ will be a constant, and it can then be taken outside the integral. In particular, k simplifies to $\left(\phi'(\phi^{-1}(y))\right) / \left(\sqrt{2\pi}\sigma\right)$ and the numerator will cancel with the denominator in $E[y|x]$ and $E[y^2|x]$, giving the same expressions as for $E[y_i|x]$ and $E[{y_i}^2|x]$ if $\eta = \sigma$. So, for linear functions, the Morris-Kanade method works fine, but, for nonlinear functions, it will give biased estimates, especially for functions with very large second derivatives.

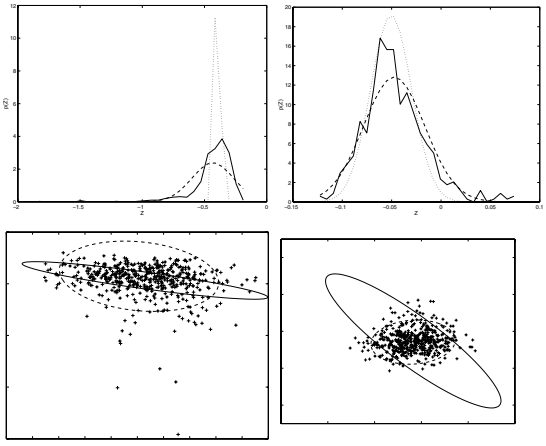

Figure 4: *Morris and Kanade's method is better than the Laplace approximation, since it allows the possibility that a distribution may not be Gaussian; however, it is not correct for nonlinear estimation functions.* **Top:** *Histograms of z-coordinates of point positions calculated by Morris and Kanade's method (solid line); Gaussian distribution calculated from mean and variance of samples (broken line); Gaussian distribution calculated from the Hessian at the mode (dotted line).* **Bottom:** *Scatterplots of points projected onto a plane through optical axis. Contour of equal probability is plotted for the Gaussian calculated from the samples (broken line) and from the Hessian (solid line). The distributions are clearly not Gaussian (they have a longer tail in one direction than the other), and furthermore, the Laplace approximation is a very poor estimate of the best Gaussian approximation.*

In figure 4, we show samples from this method applied to 25 points in 10 widely-spaced frames from the hotel data set. The samples clearly indicate a distribution with heavier tails on one side than the other. The Laplace approximation is also shown for reference, and may once again be seen to be a very poor estimate of the covariance.

2.3 Markov Chain Monte Carlo sampling

In [4], Forsyth *et al.* present a sampling approach to the structure-from-motion problem. They use a hybrid Markov Chain Monte Carlo method to draw samples from the posterior distribution of the camera parameters and point positions, given the measurements. If the chain has burnt in and the mixing rate is fast enough, the samples will have the correct covariance. As with other sampling methods, this approach has the advantage that it does not make any assumptions about the shape of the distribution.

We use a similar method here to find samples of point positions, which we have already seen are not Gaussian. In figure 1, the three sets of samples on the left are clearly non-Gaussian—their tails are much too light to be Gaussian. The three plots on the right even suggest two modes, which we cannot hope to represent with a single Gaussian.

One of the difficulties with any MCMC method is that it is usually exceedingly hard to tell if the chain has burnt in (it has forgotten its starting position) and whether it is mixing well (it is moving freely between different parts of its distribution). A correctly formulated MCMC sampler is guaranteed to converge to the desired distribution *eventually*, but this may take an impractically large number of samples. Any use of an MCMC method should be accompanied by some evidence that the samples are taken from a chain that has burnt in, and that enough samples have been generated so that we can consider the samples to be independent. Actual proofs that a chain will have a reasonable burn-in time and mixing rate are possible only for a very small number of very simple chains; usually, we must resort to some heuristics to convince ourselves that the chain has converged.

While it is possible to start the chain at a random position and have it converge to the region around the mode after a thousand samples or so, if we can start the chain close to the mode, we do not need to wait for such a long burn-in period. In our case, we start the chain at the factorisation solution, which will be reasonably close to the mode, and then reject the first hundred samples, which is probably more than necessary.

If a chain is mixing well, it will revisit the same part of its state space several times over the course of the number of samples. A plot of the trace of samples, in the order they are drawn from the chain, appears in [3]. This plot, and the fact that our chain will move to the region about the mode after sufficient samples, indicate that it is rather likely that our samples are from a chain that has burnt in, and that mixes well. This means that, after shuffling the samples, we can consider them to be independent samples from the desired distribution, in this case, the posterior on camera parameters and point positions, given the image feature positions.

3 Conclusions

It is very important to get accurate estimates of the covariance of the solution to any problem; if we draw some conclusion on the basis of the mode alone, our conclusion may often be incorrect, because the distribution actually has a very large covariance about the mode. We need to have some estimate of confidence in our solution in order to be able to make statements about it.

Even with additive Gaussian noise on the measurements, the distribution of the solutions can behave rather wildly, even exhibiting bimodal distributions on some of the marginal distributions, which we have observed in both real and toy problems. This behaviour is inherent to bilinear problems, and is not merely an artefact of the solution method or the specifics of the problem. The application of an affine transformation (*e.g.* to ensure that the camera matrix satisfies a particular model) also does not change this behaviour.

Simple methods for estimating covariances can give rather misleading estimates. Laplace's approximation assumes *a priori* that the distribution is Gaussian, which causes significant biases in the covariance estimates. Furthermore, in many cases, it does not even describe the best Gaussian approximation. The method of Morris and Kanade is better, but is still inaccurate for nonlinear functions. A properly designed and carefully used MCMC sampler will generate independent samples from the distribution, from which accurate estimates of mean and covariance may be obtained.

Acknowledgements: Research described in this paper was partially supported by NSF award 9979201.

References

[1] G. M. Crippen and T. F. Havel. *Distance geometry and molecular conformation*. Number 15 in Chemometrics Series. Research Studies Press Ltd, 1988.

[2] M. Evans and T. Swartz. *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford University Press, 2000.

[3] D.A. Forsyth, J. Haddon, and S. Ioffe. The joy of sampling. *Int. J. Computer Vision*, (In press), 2001.

[4] D.A. Forsyth, S. Ioffe, and J. Haddon. Bayesian structure from motion. In *Proc. Int. Conf. Comp. Vision*, pages 660–665, 1999.

[5] J. Koenderink and A. van Doorn. The generic bilinear calibration-estimation problem. *IJCV*, 23(3):217–234, 1997.

[6] D. Lazard. Stewart platforms and Gröbner basis. In *ARK*, pages 136–142, Ferrare, September 1992.

[7] Daniel Morris and Takeo Kanade. A unified factorization algorithm for points, line segments and planes with uncertainty models. In *Proc. Int. Conf. Comp. Vision*, pages 696–702, 1998.

[8] P.F.McLauchlan. Gauge independence in optimization algorithms for 3D vision. In *Vision Algorithms: Theory and Practice*, pages 183–198, 1999.

[9] C. Tomasi and T. Kanade. Shape and motion for image streams under orthography: a factorization method. *IJCV*, 9(2):137–154, 1992.