Factored form descent: a practical algorithm for coherence retrieval

Zhengyun Zhang,[1*] **Zhi Chen,**[1] **Shakil Rehman,**[1] **and George Barbastathis**[1,2]

[1]*Singapore-MIT Alliance for Research and Technology (SMART) Centre, 1 CREATE Way, Singapore 138602, Singapore*
[2]*Department of Mechanical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA*
[*]*zhengyun@smart.mit.edu*

Abstract: We formulate coherence retrieval, the process of recovering via intensity measurements the two-point correlation function of a partially coherent field, as a convex weighted least-squares problem and show that it can be solved with a novel iterated descent algorithm using a coherent-modes factorization of the mutual intensity. This algorithm is more memory-efficient than the standard interior point methods used to solve convex problems, and we verify its feasibility by reconstructing the mutual intensity of a Schell-model source from both simulated data and experimental measurements.

OCIS codes: (030.0030) Coherence and statistical optics; (030.4070) Modes; (100.5070) Phase retrieval.

References and links

1. M. Born and E. Wolf, *Principles of optics*, 7th. ed. (Cambridge University, 2005).
2. D. Dragoman, "Unambiguous coherence retrieval from intensity measurements," J. Opt. Soc. Am. A **20**, 290–295 (2003).
3. M. G. Raymer, M. Beck, and D. McAlister, "Complex wave-field reconstruction using phase-space tomography," Phys. Rev. Lett. **72**, 1137–1140 (1994).
4. D. F. McAlister, M. Beck, L. Clarke, A. Mayer, and M. G. Raymer, "Optical phase retrieval by phase-space tomography and fractional-order Fourier transforms," Opt. Lett. **20**, 1181–1183 (1995).
5. C. Rydberg and J. Bengtsson, "Numerical algorithm for the retrieval of spatial coherence properties of partially coherent beams from transverse intensity measurements," Opt. Express **15**, 13613–13623 (2007).
6. L. Tian, J. Lee, S. B. Oh, and G. Barbastathis, "Experimental compressive phase space tomography," Opt. Express **20**, 8296–8308 (2012).
7. B. J. Thompson and E. Wolf, "Two-beam interference with partially coherent light," J. Opt. Soc. Am. **47**, 895 (1957).
8. H. O. Bartelt, K.-H. Brenner, and A. W. Lohmann, "The Wigner distribution function and its optical production," Opt. Commun. **32**, 32–38 (1980).
9. Y. Li, G. Eichmann, and M. Conner, "Optical Wigner distribution and ambiguity function for complex signals and images," Opt. Commun. **67**, 177–179 (1988).
10. L. Waller, G. Situ, and J. W. Fleischer, "Phase-space measurement and coherence synthesis of optical beams," Nature Photon. **6**, 474–479 (2012).
11. T. Asakura, H. Fujii, and K. Murata, "Measurement of spatial coherence using speckle patterns," Optica Acta **19**, 273–290 (1972).
12. M. Michalski, E. E. Sicre, and H. J. Rabal, "Display of the complex degree of coherence due to quasi-monochromatic spatially incoherent sources," Opt. Lett. **10**, 585–587 (1985).
13. J. C. Barreiro and J. O.-C. neda, "Degree of coherence: a lensless measuring technique," Opt. Lett. **18**, 302–304 (1993).
14. H. Gamo, "Intensity matrix and degree of coherence," J. Opt. Soc. Am. **47**, 976–976 (1957).

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

15. H. M. Ozaktas, S. Yüksel, and M. A. Kutay, "Linear algebraic theory of partial coherence: discrete fields and measures of partial coherence," J. Opt. Soc. Am. A **19**, 1563–1571 (2002).
16. S. Boyd and L. Vandenberghe, *Convex Optimization* (Cambridge University , 2004).
17. E. Wolf, "New theory of partial coherence in the space-frequency domain. Part I: spectra and cross spectra of steady-state sources," J. Opt. Soc. Am. **72**, 343–351 (1982).
18. E. Polak and G. Ribiere, "Note sur la convergence de methodes de directions conjugées," Rev. Fr. Inform. Rech. O. **3**, 35–43 (1969).
19. A. C. Schell, "A technique for the determination of the radiation pattern of a partially coherent aperture," IEEE Trans. Antennas Propag. **15**, 187–188 (1967).
20. A. Papoulis, "Ambiguity function in Fourier optics," J. Opt. Soc. Am. **64**, 779–788 (1974).
21. K.-H. Brenner, A. W. Lohmann, and J. Ojeda-Castañeda, "The ambiguity function as a polar display of the OTF," Opt. Commun. **44**, 323–326 (1983).

1. Coherence retrieval

The mutual intensity function [1] for a stationary quasi-monochromatic partially coherent field contains all of the information needed to predict the time-averaged intensity at any point in the field after it has passed through any known first order optical system. Thus, its measurement enables many applications in modeling, simulation and imaging. While the mutual intensity function cannot be measured directly, its reconstruction can be posed as an inverse problem – compute the mutual intensity function from a suitable number of time-averaged intensity measurements of the field after it has passed through one or more known first order optical systems. This process of retrieving the mutual intensity of a partially coherent field from intensity measurements is known as *coherence retrieval* [2].

In this process, partially coherent light is passed through one or more known first order optical systems, and the resulting light intensities provide information about, and thus constraints on, the state of coherence of the original partially coherent field. The best known coherence retrieval methods are based on phase-space tomography [2–6], although other methods do exist, such as the direct measurement of the far field intensities of two pinholes [7], the imaging of optically-produced Wigner distribution for one-dimensional fields [8, 9], spectrogram-based methods [10] and others [11–13].

Not only are there many measurement methods for retrieving the necessary information to determine the state of coherence, there are also many different algorithms that reconstruct the state of coherence from these measurements. In this paper, we instead propose a simpler yet more versatile convex mathematical formulation and a principled solution method. Our measurement method-agnostic formulation is a constrained weighted least-squares problem based on physical first principles, and it exploits the inherent positivity of the mutual intensity. We also use this positivity in designing a practical solution method.

Unlike iterated projections algorithms [5], our formulation is convex and therefore does not suffer from potential local minima problems found when projecting onto consecutive non-convex sets. Unlike methods based on Fourier space gridding or inverse Radon transforms [3,4], we take advantage of the positivity of the mutual intensity matrix in a principled way, without the use of ad-hoc regularizers or additional projections [6]. Lastly, the flexibility of our formulation and solution method allows for every single intensity measurement to be weighted differently and removes the need for measurements to be incorporated an entire plane at a time.

We will now describe our formulation in more detail by first making some common assumptions about the partially coherent field to be measured: (a) it is quasi-monochromatic, (b) it has no evanescent components, and (c) it has negligible intensity outside a finite region in the plane.

Assumption (a) indicates that the mutual intensity function is sufficient to fully describe the partially coherent field; otherwise, the full mutual coherence function would be needed. Furthermore, this assumption places a lower bound on the wavelengths present in the field. This lower bound, along with assumption (b), imposes a spatial band-limit on the field. This band-

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

limit and assumption (c) enables accurate modeling of the continuous mutual intensity function using a finite number of samples. In other words, the partially coherent field in question can be accurately described using a mutual intensity matrix, a discretized form of the mutual intensity function where the two spatial variables are replaced by row and column indexes corresponding to spatial sample locations [14, 15]. Furthermore, let us adopt a Gaussian noise model for the intensity measurements, since it can be adapted to approximately model various sources of noise in intensity measurements, including photon shot noise, sensor read-out noise, thermal noise and quantization.

With these assumptions and models in mind, we can formulate the coherence retrieval problem as the following convex problem on mutual intensity matrices J:

Problem 1.

$$\begin{array}{ll} \text{minimize} & f(J) = \sum_{m=1}^{M} \sigma_m^{-2} \left(y_m - \mathbf{k_m}^H J \mathbf{k_m} \right)^2 \\ \text{subject to} & J \succeq 0 \end{array}$$

where:

- J is a Hermitian $N \times N$ matrix.
- N is the number of spatial samples in the mutual intensity.
- M is the number of intensity measurements.
- y_m is the m^{th} intensity measurement.
- σ_m is the standard deviation of the additive Gaussian noise source for the m^{th} intensity measurement.
- $\mathbf{k_m}$ is a vector describing propagation from the original plane where J is sought to the location where y_m is measured. Let K be the $N \times M$ matrix whose columns are $\mathbf{k_m}$; this matrix is the discretized version of the transmission function $K(P,Q)$ used for propagating the mutual intensity [1].

This formulation is a constrained least-squares problem with a quadratic merit function over the space of positive semi-definite matrices of size $N \times N$. Being convex, the problem has a single globally optimal point at best and a contiguous globally optimal set at worst. In other words, no suboptimal local minima exist for this generalized formulation of coherence retrieval.

The naive approach to this convex program is to use a generic interior point method with barrier functions [16]. However, the inner loop second-order solver requires the inversion of a Hessian, which results on the order of $O(N^6)$ operations per iteration at the worst and $O(N^4)$ operations per iteration at best. Furthermore, even with optimizations such as quasi-Newton methods, storage comparable to that of a Hessian is still required, which in this case would be of size $O(N^4)$, making such an approach not scalable for large mutual intensity matrices. Since coherence retrieval isn't a generic quadratic programming problem, one would expect that its special structure can be used in designing a simpler and less memory-intensive optimization method. This is what we propose in this paper.

In Section 2, we describe the algorithm and its theoretical justification. In Section 3, we apply the algorithm to a specific case of a Schell-model source and present numerical simulations and experimental verification.

2. Factored form descent algorithm

We start by rewriting the constrained convex problem into an unconstrained problem by exploiting the fact that all positive semi-definite matrices can be factored into the product of some

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

complex matrix and its complex conjugate transpose, with no additional constraints:

$$J = XX^H \tag{1}$$

Physically, this is equivalent to saying that partially coherent fields can be represented as incoherent ensembles of coherent fields, with each coherent field represented by a single column vector in the matrix X. Many possible factorizations are possible for any particular J; if the columns of X happen to be orthogonal, then they also form a coherent-mode decomposition of the partially coherent field [15, 17]. However, to simplify discourse in the context of this paper, we will refer to the columns of X as modes even if they are not orthogonal, and we will call the space of matrices X *modes space*.

Since any J can be factored into a product of an unconstrained matrix X and its conjugate transpose, we can convert Problem 1 into the following unconstrained quartic problem over the space of complex $N \times N$ matrices X:

Problem 2.

$$\text{minimize } \hat{f}(X) = \sum_{m=1}^{M} \sigma_m^{-2} \left(y_m - \mathbf{k_m}^H X X^H \mathbf{k_m} \right)^2$$

While there are no direct methods for solving multi-variate quartic minimization problems, the above problem can be solved using iterative methods. We propose an iterative algorithm using the nonlinear conjugate gradient method [18] to solve Problem 2. In each iteration, this algorithm aims to decrease the value of the merit function by updating the factored representation X, whose columns are the modes of the current estimate of the source mutual intensity J; we will call this algorithm the *factored form descent algorithm*:

Algorithm 1. *1. Set $X(1)$ to a random $N \times N$ complex matrix, and $S(0)$ to the zero $N \times N$ matrix.*

2. For each iteration i

(a) Compute intensity errors $\Delta_m(i) = y_m - \mathbf{k_m}^H X(i) X^H(i) \mathbf{k_m}$

(b) Set the weighted error matrix $E(i)$ to be the $M \times M$ diagonal matrix with entries $\sigma_m^{-2} \Delta_m(i)$

(c) Compute the mutual intensity space steepest descent direction $D(i) = 2KE(i)K^H$.

(d) Compute the modes space steepest descent direction $G(i) = 2D(i)X(i)$

(e) If $i = 1$ or $G(i-1) = 0$, then set $\beta(i) = 0$, otherwise use the modified Polak-Ribière formula:

$$\beta(i) = \max \left[0, \text{Re} \left\{ \frac{\langle G(i), G(i) - G(i-1) \rangle}{\|G(i-1)\|_F^2} \right\} \right]$$

(f) Compute the conjugate gradient direction $S(i) = G(i) + \beta(i) S(i-1)$

(g) Find $\alpha(i)$ that minimizes the single variable quartic polynomial $\hat{f}(X(i) + \alpha S(i))$.

(h) Update the iterate $X(i+1) = X(i) + \alpha(i) S(i)$

2.1. Algorithm behavior

Except for some rare pathological cases, the above algorithm will produce a sequence of iterates $X(i)$ such that $\hat{f}(X(i))$ approaches the globally optimal value of $\hat{f}$, thus yielding a sequence of mutual intensity matrices $X(i)X^H(i)$ such that $f(X(i)X^H(i))$ approaches the globally minimal value in Problem 1. In all cases, the following theorem applies to the above algorithm:

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

Theorem 1. *Algorithm 1 produces a sequence of iterates $X(i)$ with corresponding monotonically non-increasing merit function values $\hat{f}(X(i))$ and converges when $G(i)$ becomes zero.*

This theorem describes the overall behavior of the algorithm, including the unsurprising termination criterion, and its proof is given in Appendix A. In order to determine what value of $X(i)$ the algorithm converges to, it would be useful to determine what $G(i)$ being zero implies about $X(i)X^H(i)$ with regards to the original convex problem. To do that, let us first define an orthonormal basis for the space of $N \times N$ Hermitian matrices:

- $B_{nn} = \mathbf{u_n}(i)\mathbf{u_n}^H(i)$ for $n = 1,\ldots,N$, and
- $B_{np} = (1/2)^{(1/2)}\left(\mathbf{u_n}(i)\mathbf{u_p}^H(i) + \mathbf{u_p}(i)\mathbf{u_n}^H(i)\right)$ for $n = 1,\ldots,N$ and $p = n+1,\ldots,N$, and
- $B_{pn} = j\,(1/2)^{(1/2)}\left(\mathbf{u_n}(i)\mathbf{u_p}^H(i) - \mathbf{u_p}(i)\mathbf{u_n}^H(i)\right)$ for $n = 1,\ldots,N$ and $p = n+1,\ldots,N$.

where $\mathbf{u_n}$ for $n = 1,\ldots,N$ are the left singular vectors of $X(i)$ with monotonically non-increasing singular values. That is, if $U(i)S(i)V^H(i)$ is the singular value decomposition of $X(i)$, then $\mathbf{u_n}$ are the columns of $U(i)$ and the diagonal entries of $S(i)$ are non-increasing. Let R be the rank of $X(i)$ and let $\overline{\mathscr{B}}$ be the set of $(N-R)^2$ basis matrices B_{nn}, B_{np}, B_{np} where $R < n < p$ and let $\mathscr{B}$ be the set of the remaining basis matrices. If we consider the geometry of the convex cone formed by the set of all positive semi-definite matrices, then the boundary of this cone is the set of matrices that are also rank-deficient. For a rank-deficient $X(i)$, the basis $\overline{\mathscr{B}}$ describes a space orthogonal to the boundary of the cone and protruding from $X(i)X^H(i)$.

Using the above definitions, the following theorem specifies a relationship between the modes space steepest descent direction $G(i)$ and the mutual intensity space steepest descent direction $D(i)$:

Theorem 2. *When $G(i)$ is zero, the mutual-intensity space steepest descent direction $D(i)$ at position $X(i)X^H(i)$ is orthogonal to all the basis matrices in $\mathscr{B}$.*

Proof. We can write $G(i)$ as:

$$G(i) = \sum_{n=1}^{N} D(i)\sigma_n(i)\mathbf{u_n}(i)V^H(i) \tag{2}$$

Since we know $G(i)$ to be zero and since $V(i)$ is an orthonormal matrix, we have:

$$0 = \sum_{n=1}^{N} D(i)\sigma_n(i)\mathbf{u_n}(i) \tag{3}$$

This means that $D(i)\mathbf{u_n}(i)$ is zero for $n \leq R$. Note that any quadratic form of $D(i)$ can be written as an inner product:

$$\mathbf{u_n}^H(i)D(i)\mathbf{u_p}(i) = \left\langle \mathbf{u_n}(i)\mathbf{u_p}^H(i), D(i)\right\rangle \tag{4}$$

Therefore, for any $B \in \mathscr{B}$, $\langle D(i), B\rangle = 0$. □

In other words, once $G(i) = 0$, $D(i)$ can only point in a direction perpendicular to the boundary of the convex cone. As a special case, if $X(i)$ is full-rank, then $G(i)$ being zero would imply $D(i)$ being zero. Since convex problems only have one global minimum, this implies that $X(i)X^H(i)$ is the global minimum of Problem 1. If $X(i)$ is rank-deficient, then the following theorem applies:

Theorem 3. *If $G(i) = 0$ and $D(i)$ is a negative semi-definite matrix, then $X(i)X^H(i)$ is the global optimum of the original convex problem.*

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

In other words, if we are at rank-deficient global minimum, then $X(i)X^H(i)$ lies on the boundary of the convex cone and $D(i)$ must point away from the cone. To prove this, let us consider the set of feasible (i.e. positive semi-definite) mutual intensity matrices in a local neighborhood around $X(i)X^H(i)$. Any such $\hat{J}$ in this set can be written as:

$$\hat{J} = X(i)X^H(i) + \varepsilon S \tag{5}$$

where ε is a small but positive number and S is a Hermitian matrix. There is a constraint on S to ensure that $\hat{J}$ is positive semi-definite. To explore this constraint, let us first project S onto null-space of $X(i)X^H(i)$:

$$\hat{S} = \hat{B}^H S \hat{B} \tag{6}$$

where $\hat{B}$ is a $(N-R)\times N$ matrix consisting of column vectors $\mathbf{u_{R+1}},\ldots,\mathbf{u_N}$. Given these definitions, the following lemma applies:

Lemma 1. *$\hat{J}$ is positive semi-definite if and only if $\hat{S}$ is also positive semi-definite.*

Proof. First, it is easy to see that if $\hat{S}$ is positive semi-definite, then $\hat{J}$ is also positive semi-definite. Now, consider the case that $\hat{S}$ is not positive semi-definite. Let $\hat{\mathbf{v}}$ be the eigenvector corresponding to a negative eigenvalue of $\hat{S}$. Let $\mathbf{v} = \hat{B}\hat{\mathbf{v}}$:

$$\mathbf{v}^H \hat{J} \mathbf{v} = \hat{\mathbf{v}}^H \hat{B}^H X(i)X^H(i)\hat{B}\hat{\mathbf{v}} + \varepsilon \hat{\mathbf{v}}^H \hat{S}\hat{\mathbf{v}} \tag{7}$$

$X^H(i)\hat{B}$ has to be equal to zero because we chose $\hat{B}$ to contain only eigenvectors corresponding to the zero-valued eigenvalues in $X(i)$. Therefore,

$$\mathbf{v}^H \hat{J} \mathbf{v} = \varepsilon \hat{\mathbf{v}}^H \hat{S}\hat{\mathbf{v}} \tag{8}$$

Since $\varepsilon > 0$ and $\hat{\mathbf{v}}$ is an eigenvector corresponding to a negative eigenvalue of $\hat{S}$, $\mathbf{v}^H\hat{X}\mathbf{v} < 0$ and thus $\hat{J}$ is not positive semi-definite. Therefore, $\hat{J}$ is positive semi-definite if and only if $\hat{S}$ is also positive semi-definite. □

With Lemma 1 proven, Theorem 3 can now be proven using proof by contradiction.

Proof. (of Theorem 3) In order for $X(i)X^H(i)$ to not be a local/global minimum of Problem 1, there must exist some $\hat{J}$ such that $f(\hat{J}) < f(X(i)X^H(i))$, and thus the corresponding step direction S must be aligned with the steepest descent direction:

$$\langle S, D(i)\rangle > 0 \tag{9}$$

Since $D(i)$ is negative semi-definite, it can be written as a sum of rank-one matrices:

$$D(i) = -\sum_n \mathbf{e_n}\mathbf{e_n}^H \tag{10}$$

Furthermore, since $G(i) = 0$, $\mathbf{e_n} = \hat{B}\hat{B}^H\mathbf{e_n}$ as a consequence of Theorem 2. Therefore, we can write:

$$\begin{aligned}\langle S, D(i)\rangle &= -\left\langle S, \sum_n \mathbf{e_n}\mathbf{e_n}^H \right\rangle \\ &= -\sum_n \mathbf{e_n}^H S \mathbf{e_n} \\ &= -\sum_n \left(\hat{B}^H\mathbf{e_n}\right)^H \hat{S}\left(\hat{B}^H\mathbf{e_n}\right) < 0\end{aligned} \tag{11}$$

This contradicts the alignment requirement specified by Eq. (9). Hence, if $D(i)$ is negative semi-definite, then there exists no matrix $\hat{J}$ in the neighborhood of $X(i)X^H(i)$ that exhibits a lower merit function value, and therefore $X(i)X^H(i)$ is the global minimum. □

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

If $D(i)$ is not negative semi-definite, then $X(i)$ is a saddle point of Problem 2. However, in practice, the algorithm will rarely converge to such a point because saddle points are inherently unstable. That is, if we approach such a saddle point, the iterate will "slide" off, away from the saddle point, unless it is approaching from a pathological direction, of which there is only a set of measure zero. It is also very easy to determine whether the algorithm has actually converged to a saddle point by examining the eigenvalues of the mutual intensity space steepest descent direction $D(i)$. If $D(i)$ is not negative semi-definite, then we can continue the algorithm after nudging the current iterate by a small fraction of a matrix composing of all the positive eigenvectors of $D(i)$:

$$X(i+1) = X(i) + \varepsilon \left(\max(\lambda_1, 0)\mathbf{v_1}, \ldots, \max(\lambda_N, 0)\mathbf{v_N}\right) \tag{12}$$

where $\mathbf{v_n}$ are the eigenvectors corresponding to eigenvalues λ_n of $D(i)$ and ε is a small positive number.

2.2. Algorithm complexity

While the nonlinear conjugate gradient method makes no guarantees about the number of iterations needed, it is possible to at least determine the asymptotic computational complexity of each iteration:

1. Propagated intensity can be computed by performing a matrix-matrix multiplication $K^H X(i)$ followed by finding the element-wise magnitude square of the result and then summing across columns. This results in $O(MN^2)$ operations, dominated by the matrix multiplication.
2. The weighted error can be computed in $O(M)$ operations.
3. The mutual intensity space steepest descent direction $D(i)$ is computed from a matrix-matrix multiplication resulting in $O(N^2M)$ operations.
4. The modes space steepest descent direction is another matrix-matrix multiplication, resulting in $O(N^3)$ operations.
5. Computation of $\beta(i)$ takes $O(N^2)$ operations.
6. Computation of $S(i)$ takes $O(N^2)$ operations.
7. Computing the terms of the quartic polynomial in α requires propagation of both $X(i)$ and $S(i)$, resulting in also $O(MN^2)$ operations.
8. Updating the iterate takes $O(N^2)$ operations.

Note that if we wish to solve for a $N \times N$ mutual intensity matrix with M measurements, then we need $M \geq N^2$. Hence, the computational complexity per iteration is $O(MN^2)$ or at least $O(N^4)$. With the availability of parallel computing, the runtime can be made shorter since the expensive $O(MN^2)$ matrix-matrix multiplications can be parallelized up to MN ways, massively reducing the run-time needed.

Storage-wise, the largest matrices are the intermediate products $K^H X(i)$ and $K^H S(i)$. However, the resulting output is much smaller and these computations can be split across blocks of K, reducing the size of the intermediate product at any particular instance in time. Therefore, the absolute minimal amount of storage needed would include the storage of K itself (which is the bulk of the storage) as well as iterates $X(i)$, past steepest descent directions and some constant amount of scratch space, resulting in an asymptotic storage complexity of $O(MN)$ or $O(N^3)$; this is much more efficient than the $O(N^4)$ asymptotic storage complexity needed for interior point methods.

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

3. Example application

In order to verify that the factored form descent algorithm also works in practice, we designed a verifiable one-dimensional phase-space tomography experiment to demonstrate retrieval of the *known* mutual intensity of a Schell-model source by the algorithm. In simulation, we used an ideal source and modeled the phase-space tomography optical arrangement to obtain a sequence of "captured" images, from which we reconstructed the mutual intensity of the ideal source with very little error. We also built the entire optical arrangement, including a 2-f system to generate the Schell-model source, and managed to reconstruct a mutual intensity that reasonably approximates the ideal source.

3.1. Design

The first aspect of the design was to specify a partially coherent source for the experiment. A one-dimensional Schell-model source was chosen for its prevalence in the literature and because it is easy to build one in practice.

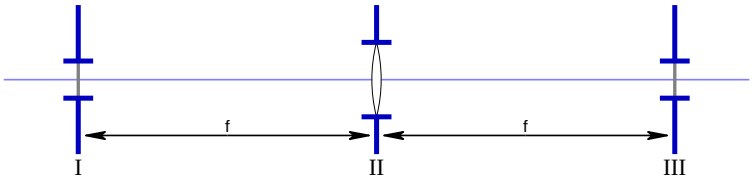

Fig. 1. Optical arrangement that generates a Schell-model beam. Uniform spatially incoherent quasi-monochromatic light is used to illuminate an amplitude mask at the front focal plane (I) of a convex lens (II) with focal length f. The partially coherent field immediately after an amplitude mask at the back focal plane (III) is that of a Schell-model source.

In general, a Schell-model source [19] in one dimension has a mutual intensity function of the form:

$$J(x_1,x_2) = a(x_1)a^*(x_2)\mu(x_1-x_2) \tag{13}$$

and can be generated using an amplitude mask illuminated by a fully incoherent area source placed effectively at infinity. An optical system consisting of an amplitude mask at both the front and back focal planes of a thin convex lens, as shown in Fig. 1, generates a Schell-model source at the back-focal plane if the front-focal plane is illuminated with uniform, fully incoherent quasi-monochromatic light. In this case, slits are used for the two masks and the resulting mutual intensity function at the back-focal plane is given by:

$$J(x_1,x_2) = I_0 \operatorname{rect}(x_1/W_2)\operatorname{rect}(x_2/W_2)\operatorname{sinc}(W_1(x_1-x_2)/(\lambda F)) \tag{14}$$

where I_0 is the maximum point-wise intensity of the output field, W_1 is the width of the front-focal plane slit, W_2 is the width of the back-focal plane slit, F is the focal length of the thin lens and λ is the wavelength of the incoming light. Based on the availability of specific optical components, the following parameters were chosen for both the simulation and the actual experiment:

$$\lambda = 532\text{nm}\,, \quad F = 100\text{mm}\,, \quad W_1 = 100\mu\text{m}\,, \quad W_2 = 500\mu\text{m} \tag{15}$$

As was discussed earlier, there are many capture methods to obtain the data needed to reconstruct the source mutual intensity. The most prevalent methods are phase-space tomography methods, where the transverse intensity of a partially coherent beam is captured at various axial

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

positions along the beam, i.e. a focal stack. The idea is that Fourier transforms of the intensity form different slices through the origin of the ambiguity function [20, 21], which in turn can be mapped one-to-one to the mutual intensity.

However, if we consider the ambiguity function where the horizontal axis is spatial frequency and the vertical axis is spatial distance, then a simple focal stack will only capture a fraction of the first and third quadrants. In fact, the horizontal axis (corresponding to the plane where we wish to obtain the mutual intensity) and the vertical axis (corresponding to a plane infinitely far away) cannot be measured.

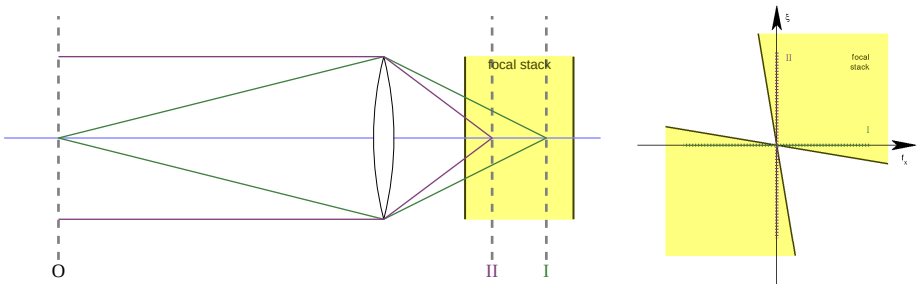

Fig. 2. Using a convex lens, as shown in (a), we can capture both the image (I) and the Fourier transform intensity (II) of a partially coherent source located at (O) using a finitely deep focal stack. That is, the intensity at planes I and II correspond to the horizontal and vertical axes respectively of the ambiguity function (b), allowing more of the ambiguity function to be directly measured compared to a lensless approach.

Instead, we propose a slightly modified arrangement wherein we use a convex lens to map both the image (horizontal axis) and the Fourier plane (vertical axis) to locations we can image directly using a sensor, as shown in Fig. 2. Thus, we can capture the entire first and third quadrants along with some parts of the second and fourth quadrants.

With this modification, we also bypass another problem of standard focal stacks, where the light diffracts outward and eventually requires a large and very sensitive sensor to capture the entire field. By using the lens, we are compressing not only the axial extent, but also the transverse extent of the propagating field so that a finite sensor moving through a finite distance obtains information about more than half the domain of the ambiguity function.

With this additional modification to phase space tomography, two extra parameters need to be decided for the design of the optics in the capture system – the focal length of the lens and the distance from the source plane to the lens. Higher values result in smaller effective numerical aperture but less aberrations, assuming fixed lens diameters. However, higher values would also result in a longer distance between the images of the Fourier and source planes. Thus, a compromise would have to be reached for each particular situation. For this particular design, we decided on using a convex lens with focal length 50mm placed 150mm after the source plane, which yields images of the Fourier and primal planes at 50mm and 75mm behind the lens respectively.

After the capture optics, all that remains are the camera and the linear translation stage. We used a camera with 3.2μm pixel pitch and modeled it as an ideal sampling system with sampling interval 3.2μm. According to Nyquist, the highest frequency pattern in intensity that could be captured would have period 6.4μm, corresponding to a 12.8μm period in the field. Since the image of the source plane is demagnified by a factor of 2 due to the convex lens, effectively the

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

highest frequency field component in the source plane we could image would have a period of 25.6μm. This corresponds to a sampling rate of 12.8μm in the field. The input source should be ideally 500μm across, but we modelled it as a source 1500μm across so as to not introduce such a strong prior into the retrieval process, since anything outside the source region is considered to have zero intensity by the algorithm. This results in 118 pixels at a sampling rate of 12.8μm for the model of the source field, and thus a 118×118 mutual intensity matrix.

The second slit had width 500μm, which corresponds to 78.125 pixels on the camera when the sensor plane was on the image of the source plane. The first slit had width 100μm, which corresponds to 15.625 pixels on the camera when the sensor plane was on the image of the Fourier plane. Thus, 256 pixels across on the sensor would be able to fully cover both planes and give enough breathing room on either side.

The last parameter to determine was how many images to capture. Since we are trying to recover a 118×118 mutual intensity matrix, we need at least 13924 independent data points for input to the algorithm. Since each camera image is 256 pixels long, we'd need at least 55 images. The stage available had a total travel of 50mm and thus we decided to capture a total of 101 images placed 0.25mm apart.

3.2. Simulation

For mutual intensity recovery from simulation, a noiseless set of intensity measurements $\mathbf{y_0}$ was obtained through simulated propagation of the partially coherent field. Propagation was computed by taking the source mutual intensity, in this case a theoretical mutual intensity computed using Eq. (14) and the design parameters in Eq. (15). $\mathbf{y_0}$ is visualized as a focal stack in Fig. 3. The gamma-boosted image is included to show that the apparent sharp bending of light at the image plane is an illusion.

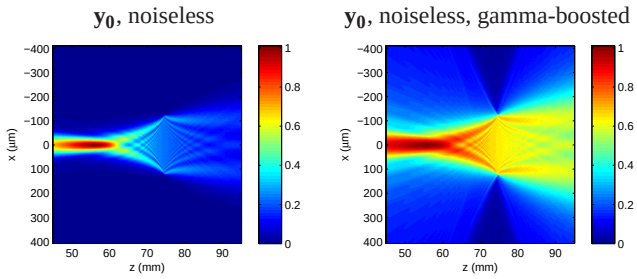

Fig. 3. Simulated noiseless data used as input to the factored form descent algorithm, visualized as focal stacks. The right image has been gamma-boosted with $\gamma = 0.3$ to enhance the visibility of lower energy parts of the stack.

From $\mathbf{y_0}$, three noisy sets of data were also generated (and they are shown as focal stacks in Fig. 4):

1. $\mathbf{y_{01}}$ – uniform Gaussian noise in intensity with standard deviation equal to 0.1% of the maximum intensity in $\mathbf{y_0}$.

2. $\mathbf{y_1}$ – uniform Gaussian noise in intensity with standard deviation equal to 1% of the maximum intensity in $\mathbf{y_0}$.

3. $\mathbf{y_p}$ – uniform Gaussian noise in intensity with standard deviation equal to 0.1% of the maximum intensity in $\mathbf{y_0}$ in addition to Poisson noise. The Poisson noise was simulated using Gaussian noise with variance proportional to the noiseless intensity such that there

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

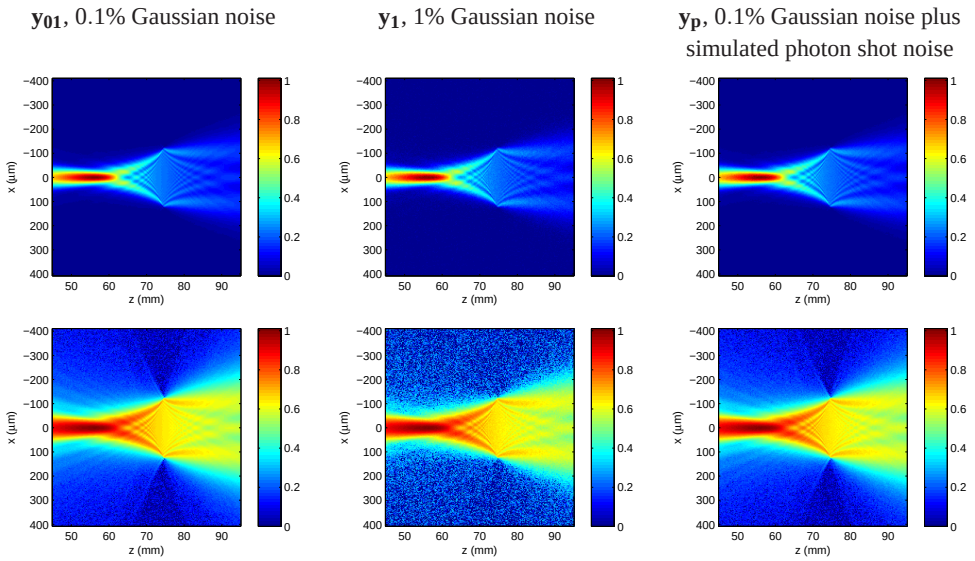

Fig. 4. Simulated noisy data used as input to the factored form descent algorithm, visualized as focal stacks (top row) and gamma-boosted focal stacks (bottom row).

was effectively Gaussian noise with standard deviation equal to 1% of the maximum intensity at the maximum intensity point. Physically, this simulation would correspond to 10000 photons at the brightest camera pixel and an average of approximately 689 photons per camera pixel across all the "captured" pixels.

These simulated data sets, including the noiseless data set, were used as input to the factored form descent algorithm. Furthermore, the Poisson noise data set $\mathbf{y_p}$ was run twice, once with uniform weighting and once with weighting that matched the varying noise standard deviations for each data point. The different runs are summarized in Table 1, with convergence of RMS intensity error shown in Fig. 5. Since we also have access to the original mutual intensity that generated the noiseless data set, we can compare the "error" between the current iterate and the theoretical mutual intensity across iterations as well. A graph of this mutual intensity error is shown in Fig. 6. Note that while the RMS intensity error is directly related to the value of the merit function $\hat{f}$ for cases of uniform weighting, it is not the case for `RUN_WP`, so the intensity RMS error may increase from one iteration to the next even if the merit function value is dropping. However, the overall progression of the merit function is essentially the same as the intensity RMS error and has been omitted from this manuscript for brevity.

From the two graphs, the following observations can be made:

- Even though the error in mutual intensity can sometimes increase, the overall progression of mutual intensity RMS error can be predicted by the overall progression of intensity RMS error.
- RMS intensity error converged for all but the noiseless data set, with the latter continuing make gains per iteration, albeit sub-linearly. It appears that noisier input sources resulted in faster "convergence".
- As expected, noisier input sources resulted in higher RMS error after convergence in both intensity and mutual intensity. For example, $\mathbf{y_1}$ showed greater error than $\mathbf{y_{01}}$ and $\mathbf{y_p}$

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

Table 1. Algorithm Runs on Simulated Data

Name	*Input*	*Iterations*	*Weighting*
RUN_0	$\mathbf{y_0}$	500	uniform
RUN_01	$\mathbf{y_{01}}$	500	uniform
RUN_1	$\mathbf{y_1}$	500	uniform
RUN_WP	$\mathbf{y_p}$	500	matching
RUN_UP	$\mathbf{y_p}$	500	uniform

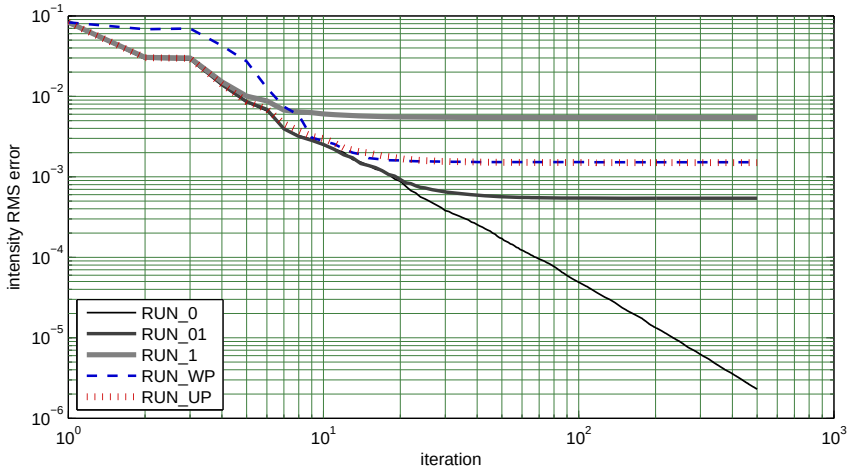

Fig. 5. Convergence of RMS error between the input intensity data set and the intensity computed from the current iterate of the mutual intensity in the factored form descent algorithm. Both the error axis and the iteration (time) axis are shown in log scale.

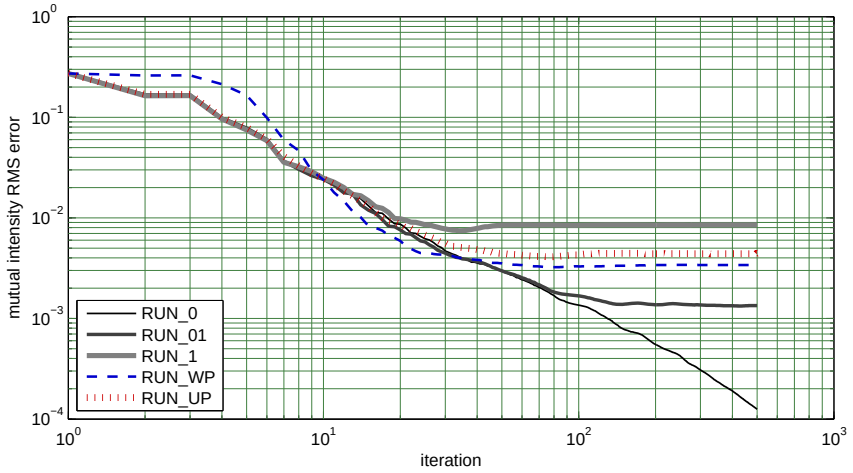

Fig. 6. Convergence of RMS error between the theoretical mutual intensity and the current iterate of the mutual intensity. Both the error axis and the iteration (time) axis are shown in log scale.

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

results were somewhere in between; $\mathbf{y_p}$ also has higher noise than $\mathbf{y_{01}}$ but on average less noise than $\mathbf{y_1}$ because peak noise for the $\mathbf{y_p}$ is only slightly higher magnitude than $\mathbf{y_1}$.

- As can be seen in Fig. 6, proper weighting using noise statistics results in some gain in reconstruction fidelity, illustrating the merits of the algorithm's ability to incorporate per-measurement weighting.

Now let us examine in more detail the final mutual intensity obtained from each of the runs. Figure 7 and Fig. 8 contain images of the theoretical mutual intensity and the resultant mutual intensity for each of the runs as well as corresponding difference images. Figure 9 is a graph comparing the drop-off in mode intensity for all of the mutual intensity matrices involved, and Fig. 10 contains field magnitude plots of the first five modes for each mutual intensity. A few more observations can be made from these figures.

- There is hardly any difference between the `RUN_0` reconstruction and the original theoretical mutual intensity. The `RUN_01` noise data yielded slightly noticeable differences, while `RUN_1` resulted in the largest error in the mutual intensity. For the Poisson shot noise simulated data, reconstructions were better than `RUN_1` but worse than `RUN_01`. Furthermore, using only uniform weighting for the algorithm resulted in a "grainier" reconstruction, as can be seen in the `RUN_UP` result.

- The errors in the mutual intensity reconstructions seem to be concentrated in two areas: a cross-shaped section and the diagonal. The "arms" of the cross are intuitive locations for error to accumulate, because of nonzero values outside the spatial extent of the original partially coherent field. The accumulation of some error along the diagonal indicates that there's some excess energy beyond the actual modes in the reconstruction, and can be a sign of a slightly under-constrained system. That is, there may not be enough constraints to pin the global minimum onto the space of rank deficient matrices.

  In the presence of noise and uniform weighting, the noise in the mutual intensity reconstruction seems to be fairly evenly distributed. However, in the case of `RUN_WP` with simulated Poisson shot noise, the majority of the error seems to fall onto two points in the mutual intensity. Furthermore, the error for the noiseless case `RUN_0` is spread out over the mutual intensity more smoothly. More research into the shape of that error region may lead to some insight into which basis functions of the mutual intensity require more time to converge and possible ways to improve the algorithm through some sort of universal preconditioner.

- The fall-off of coherence mode intensities in the reconstructions paints a very similar picture to what has been discussed before. The mode fall-off curves of the reconstructions remain close to the theoretical fall-off curve for longer for those reconstructions which result in less error. It's curious to note that while the noiseless run resulted in a single "plateau" in the fall-off curve, the noisy runs generally had two plateaus. It appears that the direct effect of the noise is present in the first plateau and convergence/underconstrainedness is expressed in the second, longer plateau. The latter must give rise to the diagonal error structure due to the vast number of modes present.

- Finally, in the field magnitude plots for the first five modes, it appears that extra noise energy is introduced to the lower intensity modes, siphoning away from the higher intensity modes. However, this could simply be an artifact of the singular value decomposition process. In either case, this indicates that there is a gradual degradation of reconstructed modes and that even if our measurements are too easy to reconstruct the lower intensity

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

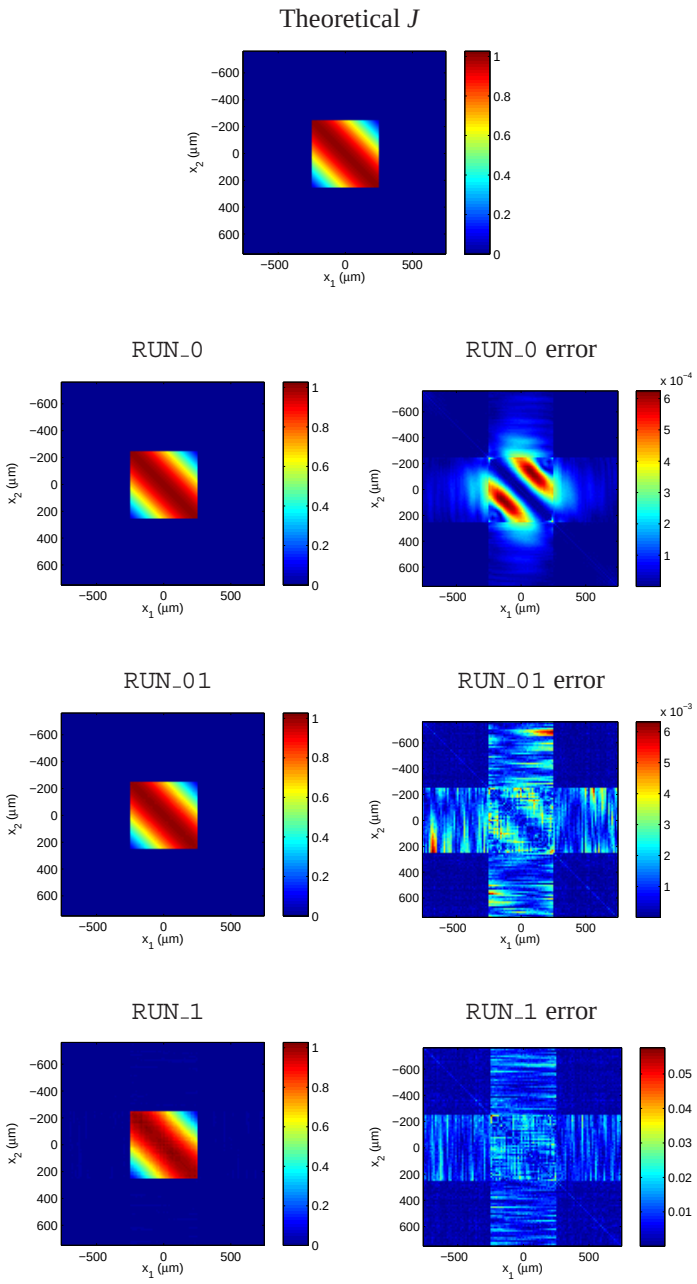

Fig. 7. At top is an image of the magnitude of the theoretical mutual intensity. Below the top image are three sets of images corresponding to three different uniform noise runs, with the left image being the magnitude of the resulting mutual intensity and the right image being the magnitude of the difference between the resulting mutual intensity and the theoretical mutual intensity.

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

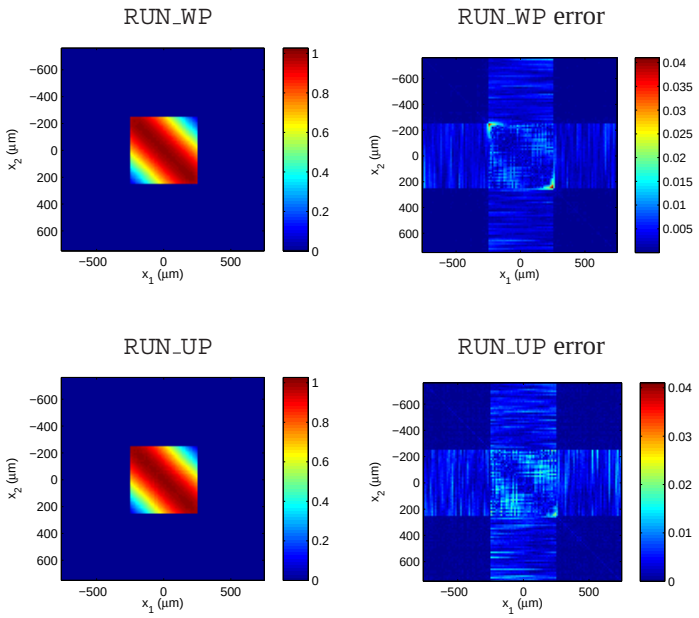

Fig. 8. Images corresponding to the runs using the Poisson shot noise data set $\mathbf{y_p}$. The left column contains images of the magnitude of the mutual intensity and the right column contains images of the magnitude of the difference between the attained mutual intensity and the theoretical mutual intensity. The error images in this case have been scaled the same to allow easier comparison between uniform weighting and matched weighting.

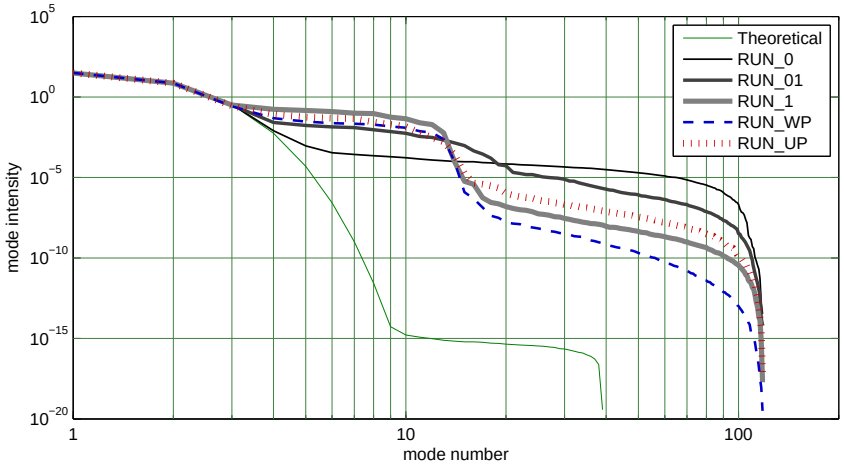

Fig. 9. A plot of the energy contained in each mode after performing a coherence mode decomposition on the theoretical mutual intensity as well as the computed mutual intensity from each run. The horizontal axis gives the mode number on a log scale, with modes sorted by decreasing energy, and the vertical axis gives the energy on a log scale.

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

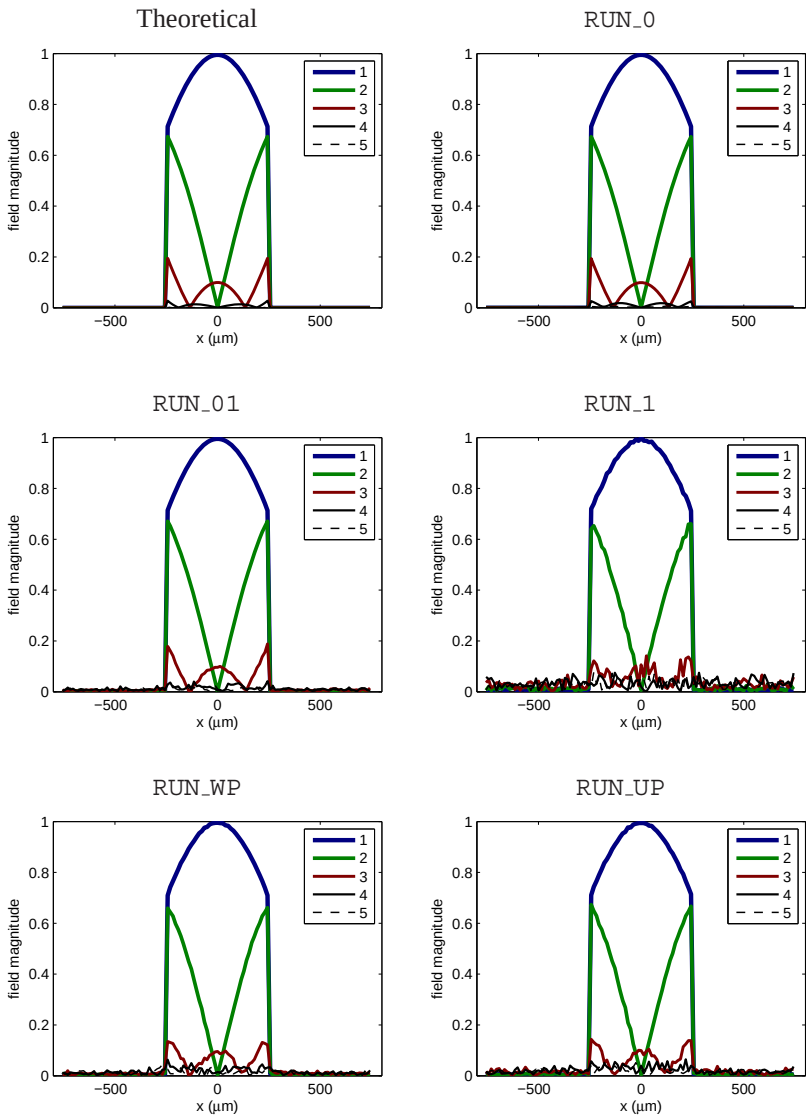

Fig. 10. A plot of the magnitude of the field for the first five coherence modes of the theoretical mutual intensity as well as the computed mutual intensity from each run.

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

modes, we can still reconstruct to a fair degree the higher intensity modes. Lastly, in comparing the uniformly weighted (RUN_UP) and properly weighted (RUN_WP) runs on the simulated Poisson shot noise data set $\mathbf{y_p}$, it is evident that using only uniform weighting results in higher noise in the modes.

3.3. Experiment

The experimental arrangement consists of two parts – the Schell-model partially coherent light source and the modified phase space tomography capture system. A Thorlabs LEDC13 530nm collimated LED light source was used as the initial light source, and its light was filtered by a 532nm bandpass filter with a FWHM of 10nm in order to enhance the monochromaticity of the light. This light was then passed through a 2-f optical system consisting of a 100mm focal length plano-convex lens with a 100nm slit at the front focal plane and a 500nm slit at the back focal plane. The flat side of the plano-convex lens was facing the 100nm slit. This arrangement generates a Schell-model partially coherent field immediately after the 500nm slit.

This partially coherent field was then propagated 150mm onwards until another plano-convex lens, this time having a focal length of 50mm and its flat side facing away. A uEye UI-1460SE-C USB color camera mounted on a Newport XMS50 motorized translation stage was placed behind this lens. The translation stage with its 50mm travel distance allowed the sensor plane on the camera to travel between locations 45mm and 95mm behind the 50mm lens.

A custom Python script was used to control both the translation stage and the camera to capture focal stacks of 201 images each, at 0.25mm step intervals. Two focal stacks were captured using the setup, one with the LED turned on (LED_ON) and one with the LED turned off (LED_OFF). The latter was captured for background subtraction.

Although the camera used was a Bayer-based color camera, the capture process captured only the raw pixel values from the sensor without demosaicing. For each stage position, the following procedure was performed:

1. Retrieve a 256×256 sub-image `img_on` from the entire 2048×1536 LED_ON image.
2. Retrieve a sub-image `img_off` covering the same exact pixels from the LED_OFF image.
3. Set `line_on` to be a set of values where each value is the mean value of only the green pixels in each column of `img_on`.
4. Set `line_off` to be a set of values where each value is the mean value of only the green pixels in each column of `img_off`.
5. Set `line_on_1` to be the set of pixel values for all the green pixels in the center 256×2 portion of `line_on`.
6. Set `line_on_std` to be the sample standard deviation of green pixel values in each column of `img_on`.
7. Subtract `line_off` from `line_on` and append to intensity measurement vector $\mathbf{y_{exp}}$.
8. Subtract `line_off` from `line_on_1` and append to intensity measurement vector $\mathbf{y_{exp1}}$.
9. Append `line_on_std` to the noise standard deviation estimate vector $\boldsymbol{\sigma}_{\mathbf{exp}}$.

The collected data is shown as focal stacks in Fig. 11.

Four separate runs of the factored form descent algorithm were then performed, as shown in Table 2. The _U runs were with uniform weighting, i.e. $\sigma_m = 1$ for all m, whereas the _W

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

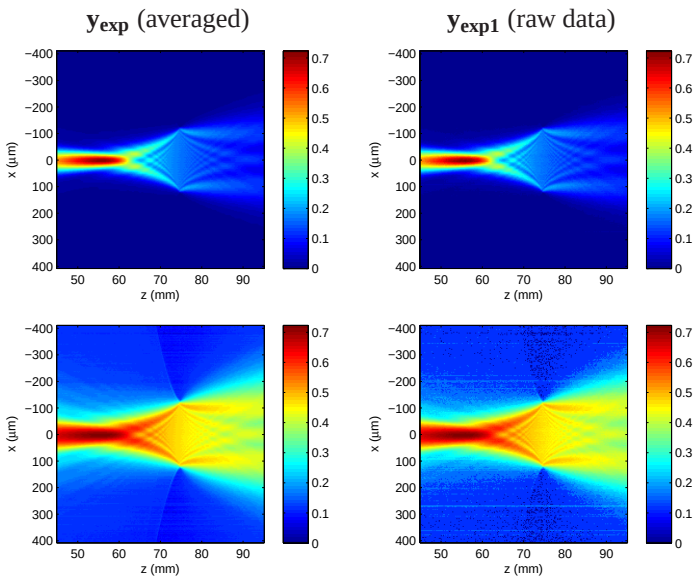

Fig. 11. The data collected during the experiment, visualized as focal stacks (top row) and gamma-boosted focal stacks (bottom row).

Table 2. Algorithm Runs on Experimental Data

Name	*Input*	*Iterations*	*Weighting*
RUN_EXP_U	$\mathbf{y_{exp}}$	500	uniform
RUN_EXP_W	$\mathbf{y_{exp}}$	500	$\sigma_{\mathbf{exp}}$
RUN_EXP1_U	$\mathbf{y_{exp1}}$	500	uniform
RUN_EXP1_W	$\mathbf{y_{exp1}}$	500	$\sigma_{\mathbf{exp}}$

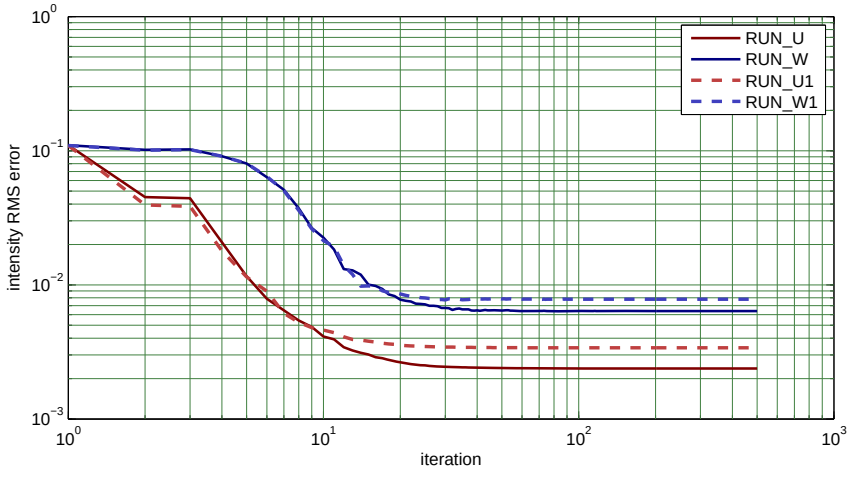

Fig. 12. Convergence of RMS error between the input experimental data and the intensity computed from the current iterate of the mutual intensity in the factored form descent algorithm. Both the error axis and the iteration (time) axis are shown in log scale.

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

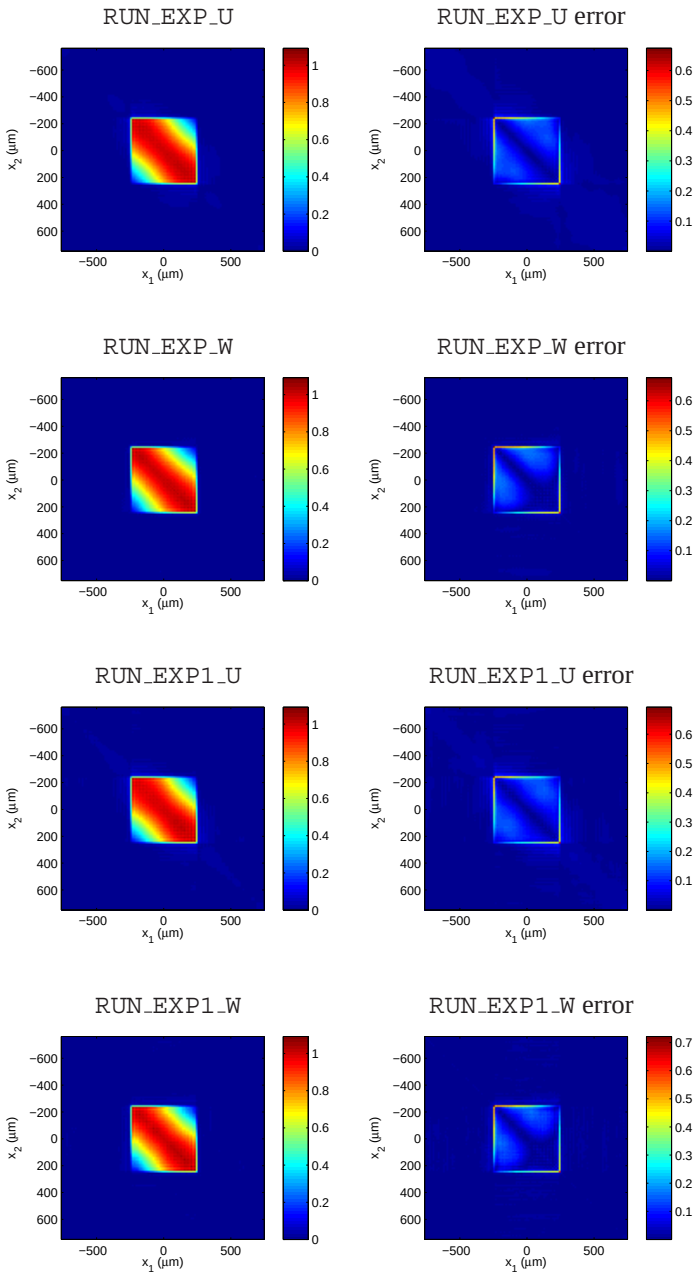

Fig. 13. Images corresponding to the resulting mutual intensity computed by runs on the experimental data sets. The left column contains images of the magnitude of the mutual intensity and the right column contains images of the magnitude of the difference between the attained mutual intensity and the theoretical mutual intensity.

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

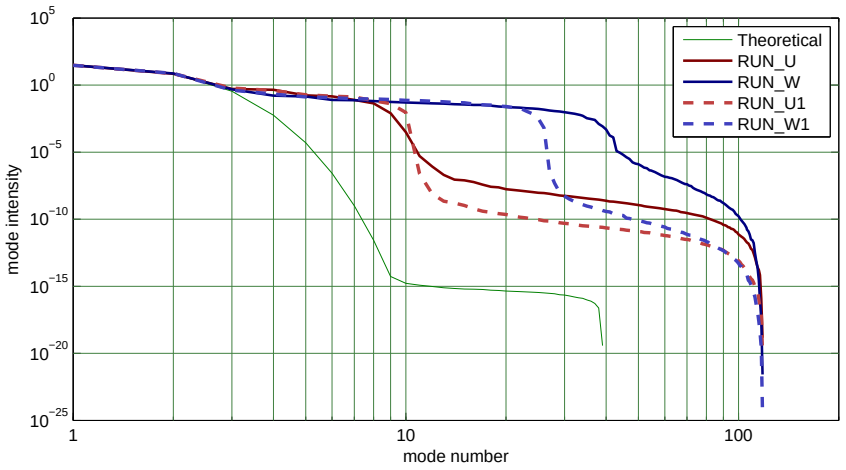

Fig. 14. A plot of the energy contained in each mode after performing a coherence mode decomposition on the theoretical mutual intensity as well as the computed mutual intensity from each experimental run. The horizontal axis gives the mode number, with modes sorted by decreasing energy, and the vertical axis gives the energy on a log scale.

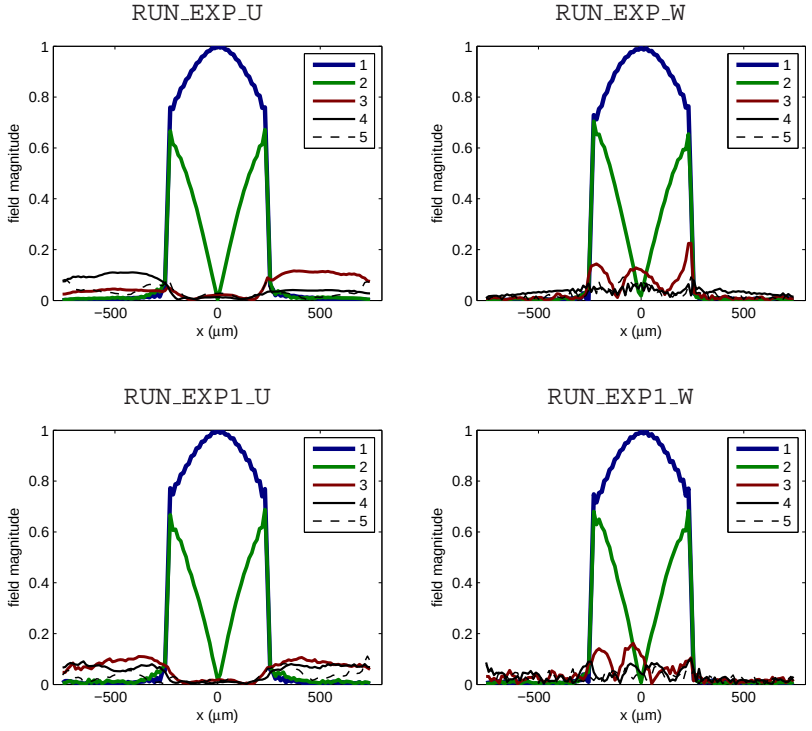

Fig. 15. A plot of the magnitude of the field for the first five coherence modes of the theoretical mutual intensity as well as the computed mutual intensity from each run.

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

runs were with estimates of the standard deviation. Though technically, the weighting should be $\sigma_{\mathbf{exp}}/\sqrt{128}$ for the RUN_EXP_W case, constant scale factors on the weighting vector have no actual effect on the algorithm; it only rescales the merit function by a constant scale factor. The RUN_EXP1 runs were done to see a more realistic use-case, since it is often not feasible to perform averages to boost SNR in real-world situations. Each run converged by the end of the 500 iterations, as can be seen in Fig. 12.

As expected, unweighted RMS intensity error was reduced in all the runs, and the weighted runs resulted in higher unweighted RMS intensity error. Furthermore, the runs on unaveraged data generally resulted in slightly higher error.

From the mutual intensity images in Fig. 13, it can be seen that the experimental data matches fairly closely the theoretical data, with most of the error in the border regions. It appears that the recovered mutual intensity wasn't as sharp as the theoretical, which is probably due to aberrations and limitations of the imaging system. Furthermore, use of noise standard deviation estimates seemed to have removed the excess energy along the diagonal in regions outside of the central square region, although the resulting mutual intensity no longer looked as spatially symmetric.

The mode-wise energy distribution graph in Fig. 14 shows that the uniformly weighted runs deviate from the theoretical mode drop-off more quickly, although they have smaller "first plateaus". The non-uniformly weighted runs follow the theoretical mode drop-off more closely, but they have larger "first plateaus" after they deviate. The higher fidelity of the first few modes can be seen in the mode field magnitudes plotted in Fig. 15, where RUN_EXP_W managed to reconstruct the first three modes, with the third mode in RUN_EXP1_W still showing some semblance of the theoretical third mode. Runs RUN_EXP_U and RUN_EXP1_U failed to reconstruct the third mode properly and exhibited more noise outside of the region of the 500nm slit.

4. Conclusions

We have constructed a generalized formulation for coherence retrieval and demonstrated a novel optimization algorithm to solve for the global minimum. A verifiable test using both simulated and experimental measurement data of a Schell-model source demonstrates that the algorithm functions correctly, leading to reasonable reconstructions of the original mutual intensity. The ability to weight the fidelity of each measurement by providing noise standard deviations for each point enables the option of boosting reconstruction fidelity by matching noise statistics. In all runs, the theoretical possibility of a pathological saddle point was never encountered.

However, there are two problems with the algorithm that could be the focus of further research. The first problem, as can be seen in Fig. 5 is that although error reduction is initially fast, final convergence is slow. Even when near convergence, there is extraneous energy present in modes that should have zero energy, as shown in Fig. 9. The noiseless case error image in Fig. 7 may provide some clues for a way to precondition the algorithm to solve these problems.

The second problem is that while the algorithm compensates for noise in the measurements, it does not compensate for errors in the specification of the propagation matrix A. In a real-world situation, it is impossible to determine the exact system function of an optical system. Error will either result from aberrations in the physical system or simply from measurement error if one chooses to characterize A empirically. To make the algorithm more robust, we need to improve the algorithm to allow it to tolerate a certain degree of error in A.

In addition to these problems, there are many other potential new avenues of research, including expanding the experimental setup to capture fully two-dimensional fields and applying the algorithm to non-tomographic capture protocols as well as illumination synthesis problems.

Lastly, we would like to discuss the similarities and differences between this work and compressive phase space tomography [6]. While both techniques aim at recovering the mutual inten-

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013

sity function of a partially coherent field and rely on the positivity of mutual intensity matrices, compressive phase space tomography introduces an additional prior that the field has a low number of modes and intentionally undersamples the measurements. This work relies on no such assumption and thus requires many more measurements. Perhaps it would be possible to unify these two approaches in the future.

A. Proof of Theorem 1

Proof. To prove Theorem 1, we will first show that the algorithm is monotonically decreasing and will eventually converge to some value. The monotonic behavior of merit function values is a direct consequence of the use of exact global line searches; for each iteration, the merit function value for the next iteration is bounded above by the current value and bounded below by the global minimum. Therefore, the sequence of merit function values must decrease monotonically and eventually converge to some value. Now let us investigate the conditions required for the algorithm to stop making progress and thus terminate.

Since the iterate $X(i)$ changes by $\alpha(i)S(i)$ each iteration, the factored form descent algorithm would stop making progress if and only if $\alpha(i)S(i)$ became zero for all future iterations. This is only possible if $S(i) = 0$ or if $\alpha(i) = 0$ in the case that $S(i) \neq 0$. We will now consider the first case.

We can split this case into two situations, depending on whether $G(i) = 0$. If $G(i) = 0$, then $\beta(i)$ necessarily has to be zero because the numerator is a dot product with $G(i)$. Therefore, $S(i) = 0$ would be zero as well. If $G(i) \neq 0$, then we can show that $S(i) \neq 0$ as well. The only way $S(i)$ could have been zero were if $G(i) = -\beta(i)S(i-1)$. For $i = 1$, $\beta(i) = 0$ and thus this statement would be false. For $i > 1$, the line search in the previous iteration guarantees that $\langle G(i), S(i-1)\rangle = 0$, and hence it would be impossible for a scalar multiple of $S(i-1)$ to be equal to $G(i)$. Therefore, $S(i) = 0$ can only be true if $G(i) = 0$.

Now let us consider the case when $S(i) \neq 0$ and $\alpha(i) = 0$. A necessary condition for $\alpha(i) = 0$ is for $G(i)$ and $S(i)$ to have an inner product of zero, necessarily implying:

$$0 = \langle S(i), G(i)\rangle = \langle G(i) + \beta(i)S(i-1), G(i)\rangle = \langle G(i), G(i)\rangle \tag{16}$$

Therefore, this case also requires $G(i) = 0$ to be true. Combining the two cases results in the conclusion that when the algorithm stops making progress and terminates, $G(i) = 0$.

Hence, the factored form descent algorithm must produce monotonically decreasing merit function values and converges when $G(i) = 0$. □

Acknowledgments

The authors would like to thank Lei Tian, Laura Waller and Jon Petruccelli for their helpful discussions. This research was supported by the National Research Foundation Singapore through the Singapore-MIT Alliance for Research and Technology's BioSyM and CENSAM research programmes.

#180864 - $15.00 USD Received 29 Nov 2012; revised 3 Feb 2013; accepted 4 Feb 2013; published 1 Mar 2013