Mon. Not. R. Astron. Soc. **317**, 965–972 (2000)

Massive lossless data compression and multiple parameter estimation from galaxy spectra

Alan F. Heavens,[1]★ Raul Jimenez[1] and Ofer Lahav[2]
[1]*Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ*
[2]*Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA*

Accepted 2000 May 15. Received 2000 April 19; in original form 1999 December 10

ABSTRACT

We present a method for radical linear compression of data sets where the data are dependent on some number M of parameters. We show that, if the noise in the data is independent of the parameters, we can form M linear combinations of the data which contain as much information about all the parameters as the entire data set, in the sense that the Fisher information matrices are identical; i.e. the method is lossless. We explore how these compressed numbers fare when the noise is dependent on the parameters, and show that the method, though not precisely lossless, increases errors by a very modest factor. The method is general, but we illustrate it with a problem for which it is well-suited: galaxy spectra, the data for which typically consist of $\sim 10^3$ fluxes, and the properties of which are set by a handful of parameters such as age, and a parametrized star formation history. The spectra are reduced to a small number of data, which are connected to the physical processes entering the problem. This data compression offers the possibility of a large increase in the speed of determining physical parameters. This is an important consideration as data sets of galaxy spectra reach 10^6 in size, and the complexity of model spectra increases. In addition to this practical advantage, the compressed data may offer a classification scheme for galaxy spectra which is based rather directly on physical processes.

Key words: methods: data analysis – methods: statistical – galaxies: fundamental parameters – galaxies: statistics.

1 INTRODUCTION

There are many instances where objects consist of many data, whose values are determined by a small number of parameters. Often, it is only these parameters which are of interest. The aim of this paper is to find linear combinations of the data which are focused on estimating the physical parameters with as small an error as possible. Such a problem is very general, and has been attacked in the case of parameter estimation in large-scale structure and the microwave background (e.g. Tegmark, Taylor & Heavens 1997, hereafter TTH; Tegmark 1997a,b; Bond, Jaffe & Knox 1998). Previous work has concentrated largely on the estimation of a single parameter; the main advance of this paper is that it sets out a method for the estimation of multiple parameters. The method provides one projection per parameter, with the consequent possibility of a massive data compression factor. Furthermore, if the noise in the data is independent of the parameters, then the method is entirely lossless, i.e. the compressed data set contains as much information about the parameters as the full data set, in the sense that the Fisher information matrix is the same for the compressed data set as the entire original data set. An equivalent statement is that the mean likelihood surface is at the peak locally identical when the full or compressed data are used.

We illustrate the method with the case of galaxy spectra, for which there are surveys underway which will provide $\sim 10^6$ objects. In this application, the noise is generally not independent of the parameters, as there is a photon shot-noise component which depends on how many photons are expected. We take a spectrum with poor signal-to-noise ratio (S/N), the noise of which is approximately from photon counting alone, and investigate how the method fares. In this case, the method is not lossless, but the increase in error bars is shown to be minimal, and superior in this respect to an alternative compression system, Principal Component Analysis (PCA).

One advantage that such radical compression offers is speed of analysis. A major scientific goal of galaxy spectral surveys is to determine physical parameters of the stellar component of the galaxies, such as the age, star formation history, initial mass function, and so on. Such a process can, in principle, be achieved by generating model galaxy spectra by stellar population synthesis techniques, and finding the best-fitting model by maximum-likelihood techniques. This can be very time-consuming, and must inevitably be automated for so many galaxies. In addition, one may have a large parameter space to explore, so any method which can speed up this process is worth investigation. One possible further

★ E-mail: afh@roe.ac.uk

application of the data compression method is that the handful of numbers might provide the basis of a classification scheme which is based on the physical properties one wants to measure.

The outline of the paper is as follows. In Section 2 we set out the lossless compression method for noise which is independent of the parameters; the proof appears in the appendix. In Section 3 we discuss the more general case where the noise covariance matrix and the mean signal both depend on the parameters. In Section 4 we show through a worked example of galaxy spectra that the method, though not lossless, works very well in the general case.

2 METHOD

We represent our data by a vector $\boldsymbol{x}_i$, $i = 1, \ldots, N$ (e.g., a set of fluxes at different wavelengths). These measurements include a signal part, which we denote by $\boldsymbol{\mu}$, and noise, $\boldsymbol{n}$:

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{n}. \tag{1}$$

Assuming that the noise has zero mean, $\langle \boldsymbol{x} \rangle = \boldsymbol{\mu}$. The signal will depend on a set of parameters $\{\theta_\alpha\}$, which we wish to determine. For galaxy spectra, the parameters may be, for example, age, magnitude of source, metallicity, and some parameters describing the star formation history. Thus $\boldsymbol{\mu}$ is a noise-free spectrum of a galaxy with certain age, metallicity, etc.

The noise properties are described by the noise covariance matrix, $\mathbf{C}$, with components $C_{ij} = \langle n_i n_j \rangle$. If the noise is Gaussian, the statistical properties of the data are determined entirely by $\boldsymbol{\mu}$ and $\mathbf{C}$. In principle, the noise can also depend on the parameters. For example, in galaxy spectra, one component of the noise will come from photon counting statistics, and the contribution of this to the noise will depend on the mean number of photons expected from the source.

The aim is to derive the parameters from the data. If we assume uniform priors for the parameters, then the a posteriori probability for the parameters is the likelihood, which for Gaussian noise is

$$\mathcal{L}(\theta_\alpha) = \frac{1}{(2\pi)^{N/2}\sqrt{\det(\mathbf{C})}} \exp\left[-\frac{1}{2}\sum_{i,j}(x_i - \mu_i)\mathbf{C}^{-1}_{ij}(x_j - \mu_j)\right]. \tag{2}$$

One approach is simply to find the (highest) peak in the likelihood, by exploring all parameter space, and using all N pixels. The position of the peak gives estimates of the parameters which are asymptotically (low noise) the best unbiased estimators (see TTH). This is therefore the best we can do. The maximum-likelihood procedure can, however, be time-consuming if N is large and the parameter space is large. The aim of this paper is to see whether we can reduce the N numbers to a smaller number, without increasing the uncertainties on the derived parameters θ_α. To be specific, we try to find a number $N' < N$ of linear combinations of the spectral data $\boldsymbol{x}$ which encompass as much as possible of the information about the physical parameters. We find that this can be done losslessly in some circumstances; the spectra can be reduced to a handful of numbers without loss of information. The speed-up in parameter estimation is about a factor $\sim$100.

In general, reducing the data set in this way will lead to larger error bars in the parameters. To assess how well the compression is doing, consider the behaviour of the (logarithm of the) likelihood function near the peak. Performing a Taylor expansion and truncating at the second-order terms,

$$\ln \mathcal{L} = \ln \mathcal{L}_{\rm peak} + \frac{1}{2}\frac{\partial^2 \ln \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta}\Delta\theta_\alpha \Delta\theta_\beta. \tag{3}$$

Truncating here assumes that the likelihood surface itself is adequately approximated by a Gaussian everywhere, not just at the maximum-likelihood point. The actual likelihood surface will vary when different data are used; on average, though, the width is set by the (inverse of the) Fisher information matrix:

$$\mathbf{F}_{\alpha\beta} \equiv -\left\langle \frac{\partial^2 \ln \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle, \tag{4}$$

where the average is over an ensemble with the same parameters but different noise.

For a single parameter, the Fisher matrix $\mathbf{F}$ is a scalar F, and the error on the parameter can be no smaller than $F^{-1/2}$. If the data depend on more than one parameter, and all the parameters have to be estimated from the data, then the error is larger. The error on one parameter α (marginalized over the others) is at least $[(\mathbf{F}^{-1})_{\alpha\alpha}]^{1/2}$ (Kendall & Stuart 1969). There is a little more discussion of the Fisher matrix in TTH. The Fisher matrix depends on the signal and noise terms in the following way (TTH, equation 15):

$$\mathbf{F}_{\alpha\beta} = \tfrac{1}{2}{\rm Tr}[\mathbf{C}^{-1}\mathbf{C}_{,\alpha}\mathbf{C}^{-1}\mathbf{C}_{,\beta} + \mathbf{C}^{-1}(\boldsymbol{\mu}_{,\alpha}\boldsymbol{\mu}^{\rm t}_{,\beta} + \boldsymbol{\mu}_{,\beta}\boldsymbol{\mu}^{\rm t}_{,\alpha})], \tag{5}$$

where the comma indicates derivative with respect to the parameter. If we use the full data set $\boldsymbol{x}$, then this Fisher matrix represents the best that can possibly be done via likelihood methods with the data.

In practice, some of the data may tell us very little about the parameters, either through being very noisy, or through having no sensitivity to the parameters. So, in principle we may be able to throw some data away without losing very much information about the parameters. Rather than throwing individual data away, we can do better by forming linear combinations of the data, and then throwing away the combinations which tell us least. To proceed, we first consider a single linear combination of the data:

$$y \equiv \boldsymbol{b}^{\rm t}\boldsymbol{x} \tag{6}$$

for some weighting vector $\boldsymbol{b}$ (t indicates transpose). We will try to find a weighting which captures as much information about a particular parameter, θ_α. If we assume that we know all the other parameters, this amounts to maximizing $\mathbf{F}_{\alpha\alpha}$. The data set (now consisting of a single number) has a Fisher matrix, which is given in TTH (equation 25) by

$$\mathbf{F}_{\alpha\beta} = \frac{1}{2}\left(\frac{\boldsymbol{b}^{\rm t}\mathbf{C}_{,\alpha}\boldsymbol{b}}{\boldsymbol{b}^{\rm t}\mathbf{C}\boldsymbol{b}}\right)\left(\frac{\boldsymbol{b}^{\rm t}\mathbf{C}_{,\beta}\boldsymbol{b}}{\boldsymbol{b}^{\rm t}\mathbf{C}\boldsymbol{b}}\right) + \frac{(\boldsymbol{b}^{\rm t}\boldsymbol{\mu}_{,\alpha})(\boldsymbol{b}^{\rm t}\boldsymbol{\mu}_{,\beta})}{(\boldsymbol{b}^{\rm t}\mathbf{C}\boldsymbol{b})}. \tag{7}$$

Note that the denominators are simply numbers. It is clear from this expression that if we multiply $\boldsymbol{b}$ by a constant, we get the same $\mathbf{F}$. This makes sense: multiplying the data by a constant factor does not change the information content. We can therefore fix the normalization of $\boldsymbol{b}$ at our convenience. To simplify the denominators, we therefore maximize $\mathbf{F}_{\alpha\alpha}$ subject to the constraint

$$\boldsymbol{b}^{\rm t}\mathbf{C}\boldsymbol{b} = 1. \tag{8}$$

The most general problem has both the mean $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{C}$ depending on the parameters of the spectrum, and the resulting maximization leads to an eigenvalue problem which is non-linear in $\boldsymbol{b}$. We are unable to solve this, so we consider a case for which an analytic solution can be found. TTH showed how to solve for the case of estimation of a single parameter in two special cases: (1) when $\boldsymbol{\mu}$ is known, and (2) when $\mathbf{C}$ is known (i.e. does not depend on the parameters). We will concentrate on the latter case, but generalize to the problem of

estimating many parameters at once. For a single parameter, TTH showed that the entire data set could be reduced to a single number, with no loss of information about the parameter. We show below that, if we have M parameters to estimate, then we can reduce the data set to M numbers. These M numbers contain just as much information as the original data set; i.e. the data compression is lossless.

We consider the parameters in turn. With $\mathbf{C}$ independent of the parameters, $\mathbf{F}$ simplifies, and maximizing $\mathbf{F}_{11}$ subject to the constraint requires

$$\frac{\partial}{\partial b_i}(b_j\mu_{,1j}b_k\mu_{,1k} - \lambda b_j C_{jk} b_k) = 0, \tag{9}$$

where λ is a Lagrange multiplier, and we assume the summation convention ($j,k \in [1,N]$). This leads to

$$\boldsymbol{\mu}_{,1}(\boldsymbol{b}^{\rm t}\boldsymbol{\mu}_{,1}) = \lambda\mathbf{C}\mathbf{b} \tag{10}$$

with solution, properly normalized,

$$\boldsymbol{b}_1 = \frac{\mathbf{C}^{-1}\boldsymbol{\mu}_{,1}}{\sqrt{\boldsymbol{\mu}_{,1}^{\rm t}\mathbf{C}^{-1}\boldsymbol{\mu}_{,1}}}, \tag{11}$$

and our compressed datum is the single number $y_1 = \boldsymbol{b}_1^{\rm t}\boldsymbol{x}$. This solution makes sense – ignoring the unimportant denominator, the method weights high those data which are parameter-sensitive, and low those data which are noisy.

To see whether the compression is lossless, we compare the Fisher matrix element before and after the compression. Substitution of $\boldsymbol{b}_1$ into (7) gives

$$\mathbf{F}_{11} = \boldsymbol{\mu}_{,1}^{\rm t}\mathbf{C}^{-1}\boldsymbol{\mu}_{,1}, \tag{12}$$

which is identical to the Fisher matrix element using the full data (equation 5) if $\mathbf{C}$ is independent of θ_1. Hence, as claimed by TTH, the compression from the *entire* data set to the single number y_1 loses no information about θ_1. For example, if $\boldsymbol{\mu} \propto \theta$, then $y_1 = \sum_i \boldsymbol{x}_i / \sum_i \boldsymbol{\mu}_i$ and is simply an estimate of the parameter itself.

2.0.1 *Fiducial model*

It is important to note that y_1 contains as much information about θ_1 only if all other parameters are known, and also provided that the covariance matrix and the derivative of the mean in (11) are those at the maximum-likelihood point. We turn to the first of these restrictions in the next section, and discuss the second one here.

In practice, one does not know beforehand what the true solution is, so one has to make an initial guess for the parameters. This guess we refer to as the fiducial model. We compute the covariance matrix $\mathbf{C}$ and the gradient of the mean ($\mu_{,\alpha}$) for this fiducial model, to construct $\boldsymbol{b}_1$. The Fisher matrix for the compressed datum is (12), but with the fiducial values inserted. In general, this is not the same as Fisher matrix at the true solution. In practice, one can iterate: choose a fiducial model; use it to estimate the parameters, and then repeat, using the estimate as the estimated parameters as the fiducial model. As our example in Section 4 shows, such iteration may be completely unnecessary.

2.1 Estimation of many parameters

The problem of estimating a single parameter from a set of data is unusual in practice. Normally one has several parameters to estimate simultaneously, and this introduces substantial complications into the analysis. How can we generalize the single-parameter estimate above to the case of many parameters? We proceed by finding a second number $y_2 \equiv \boldsymbol{b}_2^{\rm t}\boldsymbol{x}$ by the following requirements:

(1) y_2 is uncorrelated with y_1. This demands that $\boldsymbol{b}_2^{\rm t}\mathbf{C}\mathbf{b}_1 = 0$.

(2) y_2 captures as much information as possible about the second parameter θ_2.

This requires two Lagrange multipliers (we normalize $\boldsymbol{b}_2$ by demanding that $\boldsymbol{b}_2^{\rm t}\mathbf{C}\mathbf{b}_2 = 1$ as before). Maximizing and applying the constraints gives the solution

$$\boldsymbol{b}_2 = \frac{\mathbf{C}^{-1}\boldsymbol{\mu}_{,2} - (\boldsymbol{\mu}_{,2}^{\rm t}\boldsymbol{b}_1)\boldsymbol{b}_1}{\sqrt{\boldsymbol{\mu}_{,2}\mathbf{C}^{-1}\boldsymbol{\mu}_{,2} - (\boldsymbol{\mu}_{,2}^{\rm t}\boldsymbol{b}_1)^2}}. \tag{13}$$

This is readily generalized to any number M of parameters. There are then M orthogonal vectors $\boldsymbol{b}_m$, $m = 1,\ldots M$, each y_m capturing as much information about parameter α_m which is not already contained in y_q; $q < m$. The constrained maximization gives

$$\boldsymbol{b}_m = \frac{\mathbf{C}^{-1}\boldsymbol{\mu}_{,m} - \sum_{q=1}^{m-1}(\boldsymbol{\mu}_{,m}^{\rm t}\boldsymbol{b}_q)\boldsymbol{b}_q}{\sqrt{\boldsymbol{\mu}_{,m}\mathbf{C}^{-1}\boldsymbol{\mu}_{,m} - \sum_{q=1}^{m-1}(\boldsymbol{\mu}_{,m}^{\rm t}\boldsymbol{b}_q)^2}}. \tag{14}$$

This procedure is analogous to Gram–Schmidt orthogonalization with a curved metric, with $\mathbf{C}$ playing the role of the metric tensor. Note that the procedure gives precisely M eigenvectors and hence M numbers, so the data set has been compressed from the original N data down to the number of parameters M.

Since, by construction, the numbers y_m are uncorrelated, the likelihood of the parameters is obtained by multiplication of the likelihoods obtained from each statistic y_m. The y_m have mean $\langle y_m\rangle = \boldsymbol{b}_m^{\rm t}\boldsymbol{\mu}$ and unit variance, so the likelihood from the compressed data is simply

$$\ln\mathcal{L}(\theta_\alpha) = \text{constant} - \sum_{m=1}^{M}\frac{(y_m - \langle y_m\rangle)^2}{2}, \tag{15}$$

and the Fisher matrix of the combined numbers is just the sum of the individual Fisher matrices. Note once again the role of the fiducial model in setting the weightings $\boldsymbol{b}_m$: the orthonormality of the new numbers holds only if the fiducial model is correct. Multiplication of the likelihoods is thus only approximately correct, but iteration could be used if desired.

2.1.1 *Proof that the method can be lossless for many parameters*

Under the assumption that the covariance matrix is independent of the parameters, reduction of the original data to the M numbers y_m results in no loss of information about the M parameters at all. In fact, the set $\{y_m\}$ produces, on average, a likelihood surface which is locally identical to that from the entire data set – no information about the parameters is lost in the compression process. With the restriction that the information is defined locally by the Fisher matrix, the set $\{y_m\}$ is a set of sufficient statistics for the parameters $\{\theta_\alpha\}$ (e.g. Koch 1999). A proof of this for an arbitrary number of parameters is given in the appendix.

3 THE GENERAL CASE

In general, the covariance matrix does depend on the parameters, and this is the case for galaxy spectra, where at least one component of the noise is parameter-dependent. This is the photon counting noise, for which $\mathbf{C}_{ii} = \boldsymbol{\mu}_i$. TTH argued that it is better to treat this case by using the n eigenvectors which arise from assuming the mean is known, rather than the single number (for one parameter) which arises if we assume that the covariance matrix is known, as above. We find that, on the contrary, the small number of eigenvectors $\boldsymbol{b}_m$ allow a much greater degree of compression than the known-mean eigenvectors (which in this case are simply individual pixels, ordered by $|\boldsymbol{\mu}_{,\alpha}/\boldsymbol{\mu}|$). For data signal-to-noise ratios of around 2, the latter allow a data compression by about a factor of 2 before the errors on the parameters increase substantially, whereas the method here allows drastic compression from thousands of numbers to a handful. To show what can be achieved, we use a set of simulated galaxy spectra to constrain a few parameters characterizing the galaxy star formation history.

3.1 Parameter eigenvectors

In the case when the covariance matrix is independent of the parameters, it does not matter which parameter we choose to form y_1, y_2, etc., as the likelihood surface from the compressed numbers is, on average, locally identical to that from the full data set. However, in the general case, the procedure does lose information, and the amount of information lost could depend on the order of assignment of parameters to m. If the parameter estimates are correlated, as we will see in Fig. 2, the error in both parameters is dominated by the length of the likelihood contours along the 'ridge'. It makes sense then to diagonalize the matrix of second derivatives of $\ln \mathcal{L}$ at the fiducial model, and use these as the parameters (temporarily), as proposed by Ballinger et al. (in preparation) for galaxy surveys. The parameter eigenvalues would order the importance of the parameter combinations to the likelihood. The procedure would be to take the smallest eigenvalue (with eigenvector lying along the ridge), and make the likelihood surface as narrow as possible in that direction. One then repeats along the parameter eigenvectors in increasing order of eigenvalue.

Specifically, diagonalize $\mathbf{F}_{\alpha\beta}$ in (5), to form a diagonal covariance matrix $\boldsymbol{\Lambda} = \mathbf{S}^{\mathrm{t}}\mathbf{F}\mathbf{S}$. The orthogonal parameter combinations are $\psi = \mathbf{S}^{\mathrm{t}}\theta$, where $\mathbf{S}$ has the normalized eigenvectors of $\mathbf{F}$ as its columns. The weighting vectors $\boldsymbol{b}_m$ are then computed from (14) by replacing $\boldsymbol{\mu}_{,\alpha p}$ by $\mathbf{S}_{pr}\boldsymbol{\mu}_{,\alpha r}$.

4 A WORKED EXAMPLE: GALAXY SPECTRA

We start by investigating a two-parameter model. We have run a grid of stellar evolution models, with a burst of star formation at time $-t$, where $t = 0$ is the present day. The star formation rate is $SFR(t') = A\delta(t' + t)$, where δ is a Dirac delta function. The two parameters to determine are age t and normalization A. Fig. 1 shows some spectra with fixed normalization ($1\,\mathrm{M}_\odot$ of stars produced) and different age. There are $n = 352$ pixels between 300 and 1000 nm. Real data will be more complicated (variable transmission, instrumental noise, etc.), but this system is sufficiently complex to test the methods in essential respects. For simplicity, we assume that the noise is Gaussian, with a variance given by the mean, $\mathbf{C} = \mathrm{diag}(\boldsymbol{\mu}_1, \ldots)$. This is appropriate for photon number counts when the number is large. We assume the same behaviour, even with small numbers, for illustration, but there is no reason why a more complicated noise model cannot be treated. It should be stressed that this is a more severe test of the model than a typical galaxy spectrum, where the noise is likely to be dominated by sources independent of the galaxy, such as CCD read-out noise or sky background counts. In the latter case, the compression method will do even better than the example here.

The simulated galaxy spectrum is one of the galaxy spectra (age 3.95 Gyr, model number 100), and the maximum signal-to-noise ratio per bin is taken to be 2. Noise is added, approximately photon noise, with a Gaussian distribution with variance equal to the number of photons in each channel (Fig. 1). Hence $\mathbf{C} = \mathrm{diag}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots)$.

The most probable values for the age and normalization (assuming uniform priors) is given by maximizing the likelihood:

$$\mathcal{L}(\mathrm{age}, \mathrm{norm}) = \frac{1}{(2\pi)^{n/2}\sqrt{\prod_i \mu_i}} \exp\left[-\frac{1}{2}\sum_i (x_i - \mu_i)^2/\mu_i\right], \tag{16}$$

where $\boldsymbol{\mu}$ depends on age and normalization. The natural logarithm $\ln \mathcal{L}$ is shown in Fig. 2. Since this uses all the data, and all the approximations hold, this is the best that can be done, given the S/N of the spectrum. To solve the eigenvalue problem for $\boldsymbol{b}$ requires an

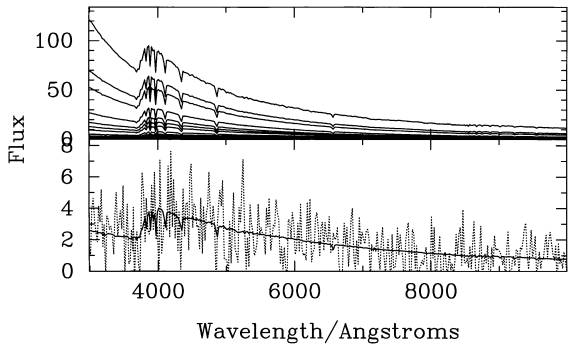

Figure 1. Top panel: example model spectra, with age increasing downwards. Bottom panel: simulated galaxy spectrum (including noise), whose properties we wish to determine, superimposed on noise-free spectrum of a galaxy with the same age.

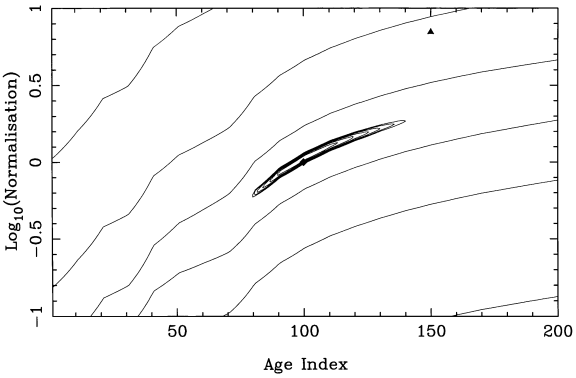

Figure 2. Full likelihood solution using all pixels. There are six contours running down from the peak value in steps of 0.5 (in $\ln \mathcal{L}$), and three outer contours at -100, -1000 and $-10\,000$. The triangle in the upper-right corner marks the fiducial model which determines the eigenvectors $\boldsymbol{b}_{1,2}$.

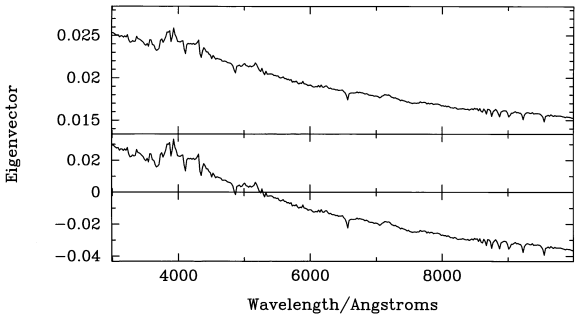

Figure 3. Eigenvectors $-\boldsymbol{b}_1$ (age) and $-\boldsymbol{b}_2$ (normalization). Wavelength λ is in Ångströms. Note that the weights in $\boldsymbol{b}_1$ are negative, which is why the sign has been changed for plotting: the blue (left) end of the spectrum which is weighted most heavily for y_1. This is expected as this part of the spectrum changes most rapidly with age. Note that these weightings differ by a constant; this feature is special to the amplitude parameter, and is explained in the text.

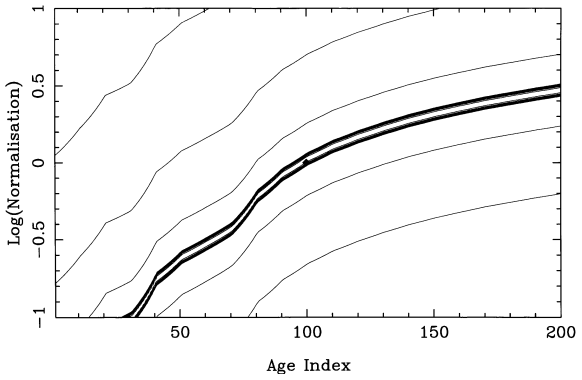

Figure 4. Likelihood solution for the age datum y_1. Contours are as in Fig. 2.

initial guess for the spectrum. This 'fiducial model' was chosen to have an age of 8.98 Gyr, i.e. very different from the true solution (model number 150 rather than 100). This allows us to compute the eigenvector $\boldsymbol{b}_1$ from (11). This gives the single number $y_1 = \boldsymbol{b}_1^{\mathrm{t}}\boldsymbol{x}$. With this as the datum, the likelihood for age and normalization is

$$\mathcal{L}(\mathrm{age}, \mathrm{norm}) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y_1 - \langle y_1 \rangle)^2}{2}\right], \tag{17}$$

where $\langle y_1 \rangle = \boldsymbol{b}_1^{\mathrm{t}}\boldsymbol{\mu}$. Note that the mean and $\langle y_1 \rangle$ here depends on the parameters – i.e. it is not from the fiducial model. The resultant likelihood is shown in Fig. 4. Clearly, it does less well than the full solution, but it does constrain the parameters to a narrow ridge, on which the true solution (age model = 100, log(normalization) = 0 lies.The second eigenvector $\boldsymbol{b}_2$ is obtained by taking the normalization as the second parameter. The vector is shown in the lower panel of Fig. 3. The normalization parameter is rather a special case, which results in $\boldsymbol{b}_2$ differing from $\boldsymbol{b}_1$ only by a constant offset in the weights. [For this parameter $\boldsymbol{\mu}_{,\alpha} = \boldsymbol{\mu}$ and so $\mathbf{C}^{-1}\boldsymbol{\mu}_{,\alpha} = (1, 1, \ldots, 1)^{\mathrm{t}}$.] The likelihood for the parameters with y_2 as the single datum is shown in Fig. 5. On its own, it does not tightly constrain the parameters, but when combined with y_1, it does remarkably well (Fig. 6).

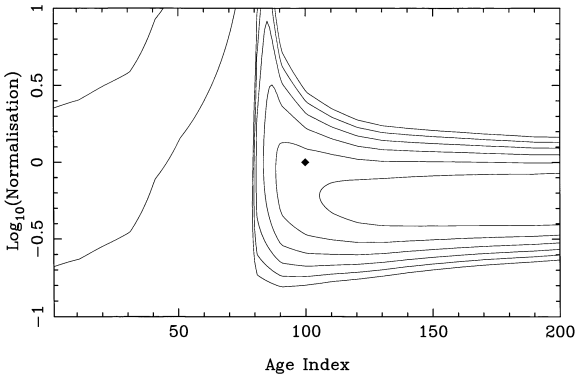

Figure 5. Likelihood solution for the normalization datum y_2. Contours are as in Fig. 2.

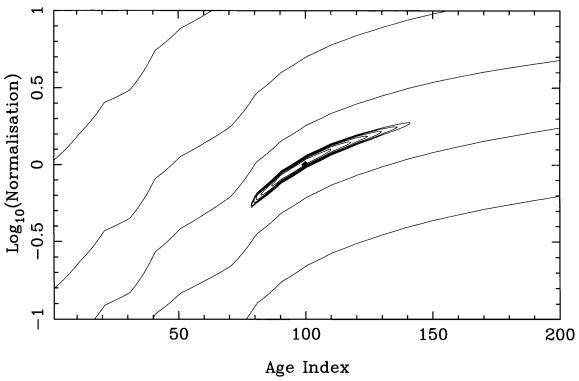

Figure 6. Likelihood solution for the age datum y_1 and the normalization datum y_2. Contours are as in Fig. 2.

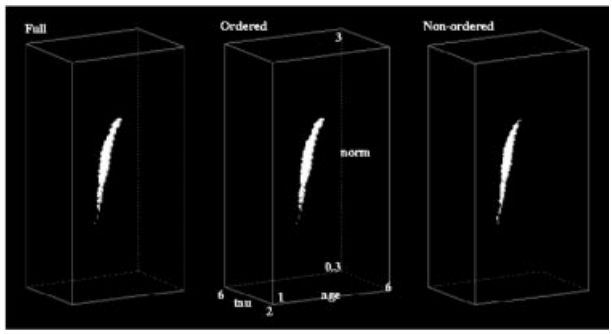

Figure 7. (Left) likelihood solution for the full data set of 1000 numbers for a single galaxy, as a function of t/Gyr, τ/Gyr and amplitude. (Middle) Likelihood for three compressed numbers, from parameter eigenvectors. (Right) likelihood surface from three compressed numbers (age, normalization and τ eigenvectors). All contours shown are 3.13 below the peak in ln $\mathcal{L}$; the irregularities in the surface are artefacts of the surface-drawing routine.

4.1 Three-parameter estimation

We complicate the situation now to a three-parameter star formation rate $SFR(t) = A\exp(-t/\tau)$, and estimate A, t and τ. Chemical evolution is included by using a simple closed-box model (with instantaneous recycling; Pagel 1997). This affects the depths of the absorption lines. If we follow the same procedure as before, choosing (t,A,τ) as the order for computing $\boldsymbol{b}_1$, $\boldsymbol{b}_2$ and $\boldsymbol{b}_3$, then the product of the likelihoods from y_1, y_2 and y_3 is as shown

in the right-hand panel of Fig. 7. The left-hand panel shows the likelihood from the full data set of 1000 numbers, which does little better than the three compressed numbers. It is interesting to explore how the parameter eigenvector method fares in this case. Here we follow the procedure in Section 2, and maximize the curvature along the ridge first. The resulting three numbers constrain the parameters as in the middle panel; in this case there is no apparent improvement over using eigenvectors from (t, A, τ), but it may be advantageous in other applications.

4.2 Estimate of increase in errors

For the noise model we have adopted, we can readily compute the increase in the conditional error for one of the parameters – the normalization of the spectrum. This serves as an illustration of how much information is lost in the compression process. In this case, $\mathbf{C} = \mathrm{diag}\,\boldsymbol{\mu}$, and $\mathbf{C}_{,\alpha} = \mathrm{diag}\,\boldsymbol{\mu}_{,\alpha} = \mathrm{diag}\,\boldsymbol{\mu}$, and the Fisher matrix (a single element) can be written in terms of the total number of photons and the number of spectral pixels. From (5), the original $F^O = N_{\mathrm{photons}} + N_{\mathrm{pixels}}/2$. The compressed data, on the other hand, have a Fisher matrix $F = N_{\mathrm{photons}} + 1/2$, so the error bar on the normalization is increased by a factor

$$\text{Fractional error increase} \simeq \sqrt{1 + \frac{1}{2s}} \qquad (18)$$

for $N_{\mathrm{photons}} \gg 1$, and $s \equiv N_{\mathrm{photons}}/N_{\mathrm{pixels}}$ is the average number of photons per pixel. Even if s is as low as 2, we see that the error bar is increased only by around 12 per cent.

4.3 Computational issues

We have reduced the likelihood problem in this case by a factor of more than 100. The eigenproblem is trivial to solve. The work to be done is in reducing a whole suite of model spectra to M numbers, and by forming scalar products of them with the vectors $\boldsymbol{b}_m$. This is a one-shot task, and trivial in comparison with the job of generating the models.

4.4 Role of fiducial model

The fiducial model sets the weightings $\boldsymbol{b}_m$. After this step, the likelihood analysis is correct for each y_m, even if the fiducial model is wrong. The only place where there is an approximation is in the multiplication of the likelihoods from all y_m to estimate finally the parameters. The y_m are strictly uncorrelated only if the fiducial model coincides with the true model. This approximation can be dropped, if desired, by computing the correlations of the y_m for each model tested. We have explored how the fiducial model affects the recovered parameters, and an example result from the two-parameter problem is shown in Fig. 8. Here the ages and normalizations of a set of 'true' galaxies with $S/N \lesssim 2$ are estimated, using a common (9-Gyr) galaxy as the fiducial model. We see that the method is successful at recovering the age, even if the fiducial model is very badly wrong. There are errors, of course, but the important aspect is whether the compressed data do significantly worse than the full data set of 352 numbers. Fig. 8 shows that this is not the case.

Although it appears from this example to be unnecessary, if one wants to improve the solution, then it is permissible to iterate, using the first estimate as the fiducial model. This adds to the computational task, but not significantly; assuming that the first iteration gives a reasonable parameter estimate, one does not have to explore the entire parameter space in subsequent iterations.

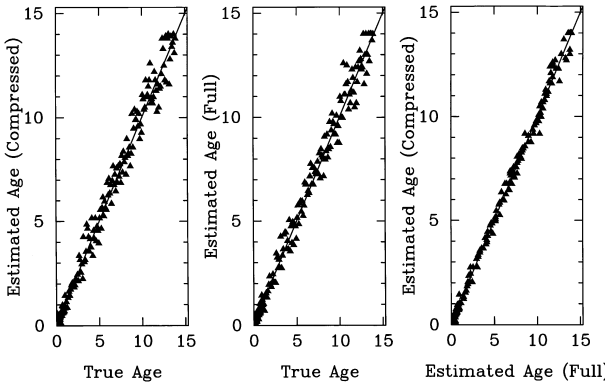

Figure 8. The effect of the fiducial model on recovery of the parameters. Here a single fiducial model is chosen (with age 9 Gyr), and ages recovered from many true galaxy spectra with ages between zero and 14 Gyr. The left-hand panel shows the recovered age from the two numbers y_1 and y_2 (with age and normalization weightings), plotted against the true model age. The middle panel shows how well the full data set (with $S/N \lesssim 2$) can recover the parameters. The right-hand panel shows the estimated age from the y_1 and y_2 plotted against the age recovered from the full data set, showing that the compression adds very little to the error, even if the fiducial model is very wrong. Note also that the scatter increases with age; old galaxies are more difficult to date accurately.

5 COMPARISON WITH PRINCIPAL COMPONENT ANALYSIS

It is interesting to compare with other data compression and parameter estimation methods. For example, Principal Component Analysis is another linear method (e.g. Murtagh & Heck 1987; Francis et al. 1992; Connolly et al. 1995; Folkes, Lahav & Maddox 1996; Sodŕe & Cuevas 1997; Bromley et al. 1998; Galaz & deLapparent 1998; Glazebrook, Offer & Deeley 1998; Singh, Gulati & Gupta 1998; Connolly & Szalay 1999; Folkes et al. 1999; Ronen, Aragon-Salamanca & Lahav 1999); this projects the data on to eigenvectors of the covariance matrix, which is determined empirically from the scatter between flux measurements of different galaxies. Part of the covariance matrix in PCA is therefore determined by differences in the models, whereas in our case $\mathbf{C}$ refers to the noise alone. PCA then finds uncorrelated projections which contribute in decreasing amounts to the variance between galaxies in the sample.

One finds that the first principal component is correlated with the galaxy age (Ronen et al. 1999). Fig. 9 shows the PCA eigenvectors obtained from a set of 20 burst model galaxies which differ only in age, and Fig. 10 shows the resultant likelihood from the first two principal components. In the language of this paper, the principal components are correlated, so the 2×2 covariance matrix is used to determine the likelihood. We see that the components do not do nearly as well as the parameter eigenvectors; they do about as well as y_1 on its own. For interest, we plot the first principal component and y_1 versus age in Fig. 11. In the presence of noise ($S/N < 2$ per bin), y_1 is almost monotonic with age, whereas PC1 is not. Since PCA is not optimized for parameter estimation, it is not lossless, and it should be no surprise that it fares less well than the tailored eigenfunctions of Section 3.

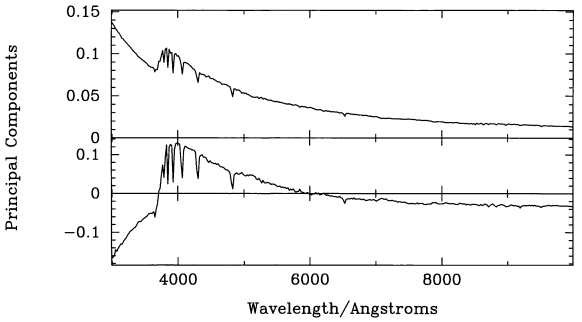

Figure 9. The first two principal component eigenvectors, from a system of model spectra consisting of a burst at different times.

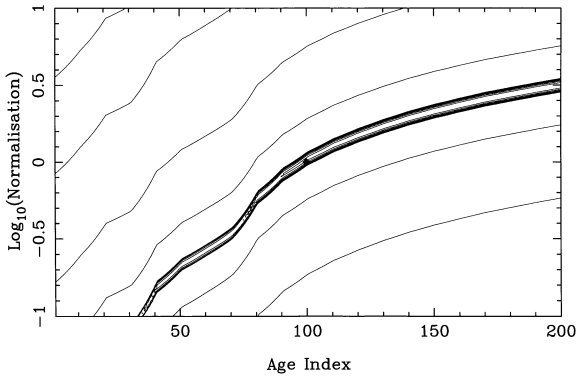

Figure 10. Likelihood solution for the first two principal components, PC1 (top) and PC2. Contours are as in Fig. 2.

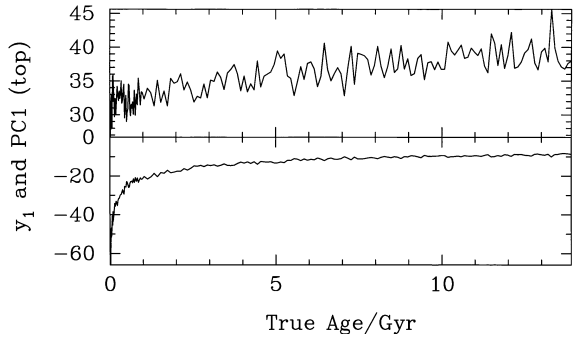

Figure 11. First principal component (PC1) and y_1 versus age. One in every 10 models was used to do the PCA. In the presence of noise, at a level of $S/N < 2$ per bin, y_1 is almost monotonic with age, whereas PC1, although correlated with age, is not a good age estimator.

If one cannot model the effect of the parameters a priori, then this method cannot be used, whereas PCA might still be an effective tool.

6 DISCUSSION

We have presented a linear data compression algorithm for estimation of multiple parameters from an arbitrary data set. If there are M parameters, the method reduces the data to a compressed data set with M members. In the case where the noise is independent of the parameters, the compression is lossless; i.e. the M data contain as much information about the parameters as the entire data set. Specifically, this means the mean likelihood surface around the peak is locally identical whichever of the full or compressed data set is used as the data. It is worth emphasising the power of this method: it is well known that, in the low-noise limit, the maximum-likelihood parameter estimates are the best unbiased estimates. Hence if we do as well with the compressed data set as with the full data set, there is no other method, linear or otherwise, which can improve upon our results. The method can result in a massive compression, with the degree of compression given by the ratio of the size of the data set to the number of parameters. Parameter estimation is speeded up by the same factor.

Although the method is lossless in certain circumstances, we believe that the data compression can still be very effective when the noise does depend on the model parameters. We have illustrated this using simulated galaxy spectra as the data, where the noise comes from photon counting (in practice, other sources of noise will also be present, and possibly dominant); we find that the algorithm is still almost lossless, with errors on the parameters increasing typically by a factor $\sim\sqrt{1+1/(2s)}$, where s is the average number of photons per spectral channel. The example we have chosen is a more severe test of the algorithm than real galaxy spectra; in reality the noise may well be dominated by factors external to the galaxy, such as detector read-out noise, sky background counts (for ground-based measurements) or zodiacal light counts (for space telescopes). In this case, the noise is indeed independent of the galaxy parameters, and the method is lossless.

The compression method requires prior choice of a fiducial model, which determines the projection vectors $\boldsymbol{b}$. The choice of fiducial model will not bias the solution, and the likelihood given the y_m individually can be computed without approximation. Combining the likelihoods by multiplication from the individual y_m is approximate, as their independence is guaranteed only if the fiducial model is correct. However, in our examples, we find that the method correctly recovers the true solution, even if the fiducial model is very different. If one is cautious, one could always iterate. There are circumstances where the choice of a good fiducial model may be more important, if the eigenvectors depend very sensitively on the model parameters. An example of this is the determination of the redshift z of the galaxy, whose observed wavelengths are increased by a factor $1+z$ by the expansion of the Universe. If the main signal for z comes from spectral lines, then the method will give great weight to certain discrete wavelengths, determined by the fiducial z. If the true redshift is different, these wavelengths will not coincide with the spectral lines. It should be stressed that the method will still allow an estimate of the parameters, including z, but the error bars will not be optimal. This may be one case where applying the method iteratively may be of great value.

We have compared the parameter estimation method with another linear compression algorithm, Principal Component Analysis. PCA is not lossless unless all principal components are used, and compares unfavourably in this respect for parameter estimation. However, one requires a theoretical model for the methods in this paper; PCA does not require one, needing instead a representative ensemble for effective use. Other, more ad hoc, schemes consider particular features in the spectrum, such as broad-band colours, or equivalent widths of lines (Worthey 1994). Each of these is a ratio of linear projections, with weightings given by the filter response or sharp filters concentrated at the line. There may well be merit in the way the weightings are constructed, but they will not in general do as well as the optimum weightings presented here. It is worth remarking on the ability of

the method to separate parameters such as age and metallicity, which often appear degenerately in some methods. In the 'external noise' case, then *provided* the degeneracy can be lifted by maximum-likelihood methods using every pixel in the spectrum, then it can also be lifted by using the reduced data. Of course, if the modelling is not adequate to estimate the parameters using all the data, then compression is not going to help at all, and one needs to think again. For example, a complication which may arise in a real galaxy spectrum is the presence of features not in the model, such as emission lines from hot gas. These can be included if the model is extended by inclusion of extra parameters. This problem exists whether the full or compressed data are used. Of course, we can use standard goodness-of-fit tests to determine whether the data are consistent with the model as specified, or whether more parameters are required.

The data compression to a handful of numbers offers the possibility of a classification scheme for galaxy spectra. This is attractive as the numbers are connected closely with the physical processes which determine the spectrum, and will be explored in a later paper. An additional realistic aim is to determine the star formation history of each individual galaxy, without making specific assumptions about the form of the star formation rate. The method in this paper provides the means to achieve this.

ACKNOWLEDGMENTS

We thank Andy Taylor and Rachel Somerville, and the referee, Paul Francis, for useful comments. Computations were made using Starlink facilities.

REFERENCES

Bond J. R., Jaffe A. H., Knox L., 1998, Phys. Rev. D, 57, 2117
Bromley B., Press W., Lin H., Kirschner R., 1998, ApJ, 505, 25
Connolly A., Szalay A., 1999, AJ, 117, 2052
Connolly A., Szalay A., Bershady M., Kinney A., Calzetti D., 1995, AJ, 110, 1071
Folkes S., Lahav O., Maddox S., 1996, MNRAS, 283, 651
Folkes S. et al., 1999, MNRAS, 308, 459
Francis P., Hewett P., Foltz C., Chaffee F., 1992, ApJ, 398, 476
Galaz G., de Lapparent V., 1998, A&A, 332, 459
Glazebrook K., Offer A., Deeley K., 1998, ApJ, 492, 98
Kendall M. G., Stuart A., 1969, The Advanced Theory of Statistics. Griffin, London
Koch K., 1999, Parameter Estimation and Hypothesis Testing in Linear Models. Springer-Verlag, Berlin
Murtagh F., Heck A., 1987, Multivariate Data Analysis. Astrophysics and Space Science Library, Reidel, Dordrecht
Pagel B., 1997, Nucleosynthesis and Chemical Evolution of Galaxies. Cambridge Univ. Press, Cambridge
Ronen R. T., Aragon-Salamanca A., Lahav O., 1999, MNRAS, 303, 284
Singh H., Gulati R., Gupta R., 1998, MNRAS, 295, 312
Sodŕe L., Cuevas H., 1997, MNRAS, 287, 137
Tegmark M., 1997a, ApJ, 480, L87
Tegmark M., 1997b, Phys. Rev. D, 55, 5895
Tegmark M., Taylor A., Heavens A., 1997, ApJ, 480, 22 (TTH)
Worthey G., 1994, ApJS, 95, 107

APPENDIX A

In this appendix, we prove that the linear compression algorithm for estimation of an arbitrary number M of parameters is lossless, provided the noise is independent of the parameters, $\mathbf{C}_{,\alpha}=0$. Specifically, loss-free means that the Fisher matrix for the set of M numbers $y_m=\boldsymbol{b}_m^{\rm t}\boldsymbol{x}$ is identical to the Fisher matrix of the original data set $\boldsymbol{x}$:

$$F^O_{\alpha\beta}=\langle\alpha|\beta\rangle\equiv\boldsymbol{\mu}^{\rm t}_{,\alpha}\mathbf{C}\boldsymbol{\mu}_{,\beta}. \tag{A1}$$

By construction, the y_m are uncorrelated, so the likelihoods multiply and the Fisher matrix for the set $\{y_m\}$ is the sum of the derivatives of the log-likelihoods from the individual y_m:

$$F_{\alpha\beta}=\sum_m F_{\alpha\beta}(m). \tag{A2}$$

From (7),

$$F_{\alpha\beta}(m)=(\boldsymbol{b}_m^{\rm t}\boldsymbol{\mu}_{,\alpha})(\boldsymbol{b}_m^{\rm t}\boldsymbol{\mu}_{,\beta}). \tag{A3}$$

With (14), we can write

$$\boldsymbol{b}_m^{\rm t}=\frac{\boldsymbol{\mu}^{\rm t}_{,m}\mathbf{C}^{-1}-\sum_{q=1}^{m-1}(\boldsymbol{b}_q^{\rm t}\boldsymbol{\mu}^{\rm t}_{,m})\boldsymbol{b}_q}{\sqrt{\langle m|m\rangle-\sum_{q=1}^{m-1}(\boldsymbol{b}_q^{\rm t}\boldsymbol{\mu}^{\rm t}_{,m})^2}}. \tag{A4}$$

Hence

$$F_{\alpha\beta}(m)=\left[\langle\alpha|m\rangle-\sum_{q=1}^{m-1}F_{\alpha m}(q)\right]\frac{\left[\langle\beta|m\rangle-\sum_{q=1}^{m-1}F_{\beta m}(q)\right]}{\left[\langle m|m\rangle-\sum_{q=1}^{m-1}F_{mm}(q)\right]} \tag{A5}$$

Consider first $\beta=m$:

$$F_{\alpha m}(m)=\langle\alpha|m\rangle-\sum_{q=1}^{m-1}F_{\alpha m}(q)$$

$$\Rightarrow F_{\alpha M}=\sum_{q=1}^{M}F_{\alpha M}(q)=\langle\alpha|M\rangle=F^O_{\alpha M} \tag{A6}$$

proving that these terms are unchanged after compression. We therefore need to consider $F_{\alpha\beta}(m)$ for α or $\beta<m$. First we note that

$$F_{\alpha\beta}(m)=\frac{F_{\alpha m}(m)F_{m\beta}(m)}{F_{mm}(m)} \tag{A7}$$

and, from (A6),

$$\sum_{q=1}^{\beta}F_{\alpha\beta}(q)=\langle\alpha|\beta\rangle. \tag{A8}$$

We want the sum to extend to M. However, the terms from $\beta+1$ to M are all zero. This can be shown as follows: (A7) shows that it is sufficient to show that $F_{\alpha m}(m)=0$ if $m>\alpha$. Setting $\beta=m$ in (A8), and reversing α and m, we get

$$\sum_{\alpha+1}^{m}F_{\alpha m}(q)=0. \tag{A9}$$

Now, the contribution from q does not depend on derivatives wrt higher-numbered parameters, so we can evaluate $F_{\alpha m}(\alpha+1)$ by setting $m=\alpha+1$. The sum (A9) implies that this term is zero. Increasing m successively by one up to M, and using (A9), proves that all the terms are zero, and hence that the compression is lossless.

This paper has been typeset from a TEX/LATEX file prepared by the author.