The Canadian Journal of Statistics
Vol. 40, No. 4, 2012, Pages 712–730
La revue canadienne de statistique

Fisher information matrix: A tool for dimension reduction, projection pursuit, independent component analysis, and more

Bruce G. LINDSAY[1] and Weixin YAO[2]*

[1] *Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA*
[2] *Department of Statistics, Kansas State University, Manhattan, KS 66502, USA*

Key words and phrases: Dimension reduction; Fisher information matrix; independent component analysis; projection pursuit; white noise matrix.

MSC 2010: Primary 62-07; secondary 62H99.

Abstract: Hui & Lindsay (2010) proposed a new dimension reduction method for multivariate data. It was based on the so-called white noise matrices derived from the Fisher information matrix. Their theory and empirical studies demonstrated that this method can detect interesting features from high-dimensional data even with a moderate sample size. The theoretical emphasis in that paper was the detection of non-normal projections. In this paper we show how to decompose the information matrix into non-negative definite information terms in a manner akin to a matrix analysis of variance. Appropriate information matrices will be identified for diagnostics for such important modern modelling techniques as independent component models, Markov dependence models, and spherical symmetry. *The Canadian Journal of Statistics* 40: 712–730; 2012 © 2012 Statistical Society of Canada

Résumé: Hui et Lindsay (2010) ont proposé une nouvelle méthode de réduction de la dimension pour les données multidimensionnelles. Elle est basée sur des matrices communément qualifiées de bruits blancs obtenues à partir de la matrice d'information de Fisher. Leurs études théoriques et empiriques montrent que cette méthode peut détecter des caractéristiques intéressantes à partir de données de grande dimension même si les échantillons sont de taille modérée. L'emphase théorique de cet article était mise sur la détection des projections non normales. Nous montrons ici comment décomposer la matrice d'information en termes d'information définie non négative de façon similaire à l'analyse de variance matricielle. Des matrices d'information appropriées peuvent être identifiées comme diagnostic pour les techniques de modélisation modernes telles que les modèles en composantes indépendantes, les modèles de dépendance markovienne et la symétrie sphérique. *La revue canadienne de statistique* 40: 712–730; 2012 © 2012 Société statistique du Canada

1. INTRODUCTION

Let $f(\mathbf{x})$ be a density function for the d-dimensional random vector $\mathbf{X}$. The density f will be treated as the unknown, and will be the fundamental object of our investigation. Let $f(\mathbf{x}-\boldsymbol{\theta})$ be the corresponding location family in $\boldsymbol{\theta}$. Assume that regularity holds, so that one can construct the $d \times d$ information matrix $J_{\mathbf{X}}$ for the parameter $\boldsymbol{\theta}$. The paper Hui & Lindsay (2010) showed that $J_{\mathbf{X}}$ can be reliably estimated and used for data-based dimension reduction to the most interesting linear combinations of a multivariate dataset. The selected projections were those that maximized information. The main purpose of this paper is to create a deeper understanding of how $J_{\mathbf{X}}$ can be decomposed into a summation of information matrices that carry information about such features of f as spherical symmetry, independence, and Markov independence.

* *Author to whom correspondence may be addressed.*
E-mail: wxyao@ksu.edu

We start with our basic motivation. With modern data gathering devices and vast data storage space, researchers can easily collect high-dimensional data, such as biotech data, financial data, satellite imagery and hyperspectral imagery. The analysis of high-dimensional data poses many challenges for statisticians. One of them is how to extract interesting features in a much reduced low-dimensional space from a high-dimensional space. Our analysis in this paper will relate $J_{\mathbf{X}}$ to such modern dimension reduction techniques as independent component analysis and graphical models, extending the projection pursuit discussion of Hui & Lindsay (2010).

Friedman & Tukey (1974) developed projection pursuit to explore high-dimensional data by examining the marginal distributions of low-dimensional linear projections. As argued later by Huber (1985) and Jones & Sibson (1987), the Gaussian distribution is the least interesting one, and so the most interesting directions should be those that show the least normality. The overall aim of projection pursuit was to find the least normal projections directly and then use them as the data summaries. One drawback of these methods has been their high computational cost, especially for high-dimensional data, due to the need to search through the large number of possible projections.

Hui & Lindsay (2010) proposed a novel dimension reduction approach to projection pursuit that changed the computational search problem into an explicit eigenanalysis problem. From the data reduction point of view, they stated that the goal was not just to find interesting directions, but also to eliminate uninteresting ones. By uninteresting they meant the projections that were most similar to "white noise"; that is, their distribution was marginally normal and also independent of all orthogonal projections. Unlike conventional projection pursuit, which is focused only on marginal features, white noise analysis included the consideration of dependence relationships. The projections used in their analysis also had an interpretation as being the most informative components, in the Fisher information sense.

To illustrate the logic of this methodology, in Figure 1 we show two plots derived from an analysis of Fisher's Iris data, which contains 150 observations of four measurements: sepal length, sepal width, petal length, and petal width. There are three classes of 50 observations each, where each class refers to a species of iris plant. As will be seen, one species is quite different from the others. We will ignore the class labels and use the white noise matrix analysis to check whether the information matrix uncovers information about the class labels. The white noise matrix is 4×4, so there are four eigenvalues and eigenvectors. We use the eigenvectors to form the projections.

In Figure 1a, we show the two projections of the data that are the most similar to white noise; they clearly do not reveal any interesting cluster information. Indeed, the plot looks like it could be multivariate normal data. However, one could propose other hypotheses about this scatterplot. Even if not normal, is the distribution spherically symmetric? Or are the two coordinate variables independent? As we will see, one can address these issues using our information decomposition.

By comparison, Figure 1b shows the scatter plot of the two *most* informative orthogonal projections. The first component, placed on the x axis, clearly separates the three species. The second component displays a more subtle effect, in that there is dependence between the first and second coordinates reflected in the change of variance across the plot. Indeed, one of the characteristics of white noise analysis is that large information values can either reflect failure of marginal normality or dependence between variables.

The emphasis in Hui & Lindsay (2010) was strictly on diagnostics for non-normality. We review these results in Section 2. However, the normal density is the unique distribution that simultaneously displays spherical symmetry and independent coordinates. In Section 3, we will show how one can create an additive decomposition of the information matrix whose components can reveal more about the multivariate structure of f, including dependence and symmetry relationships. In particular, these information components are diagnostic for such important modern modelling techniques as independent component models, graphical models, and spherical symmetry.

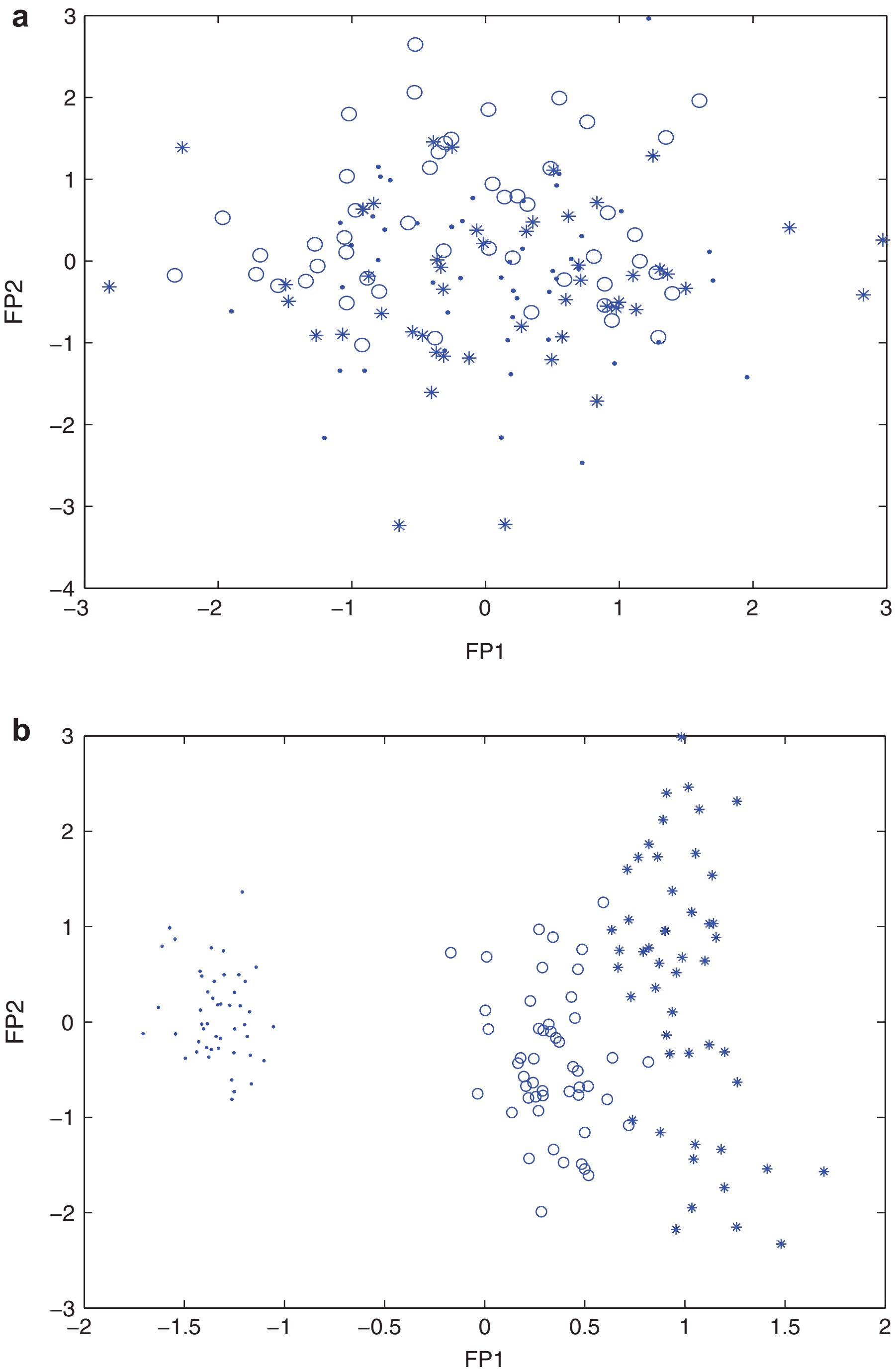

FIGURE 1: (a) The scatter plot of the two LEAST interesting coordinate projections based on white noise matrix analysis for Irish Flower data. (b) The scatter plot of the two MOST informative solution projections based on white noise matrix analysis for Irish Flower data. [Colour figure can be viewed in the online issue, which is available at wileyonlinelibrary.com]

The decomposition results of Section 3 treat f as known. In Section 4 we consider the consequences for our theory when one estimates f using a kernel density estimator. Finally, we provide some discussion in Section 5, including our thoughts on the suitability of these methods for "small n, large d" analysis. Our focus throughout the paper will be on the theoretical developments, leaving data analysis for future work.

2. WHITE NOISE MATRIX

2.1. Introduction

Hui & Lindsay (2010) developed the following approach to extract lower dimensional projections of data that contain interesting nonlinear features. It could be called *Most Informative Component Analysis*.

We find it helpful to describe the analysis as occurring in two stages of "whitening the data." We start with vector data $\mathbf{X}$. In the first stage, we standardize the data to the vector $\mathbf{Y} = \Sigma^{-1/2}\mathbf{X}$, where $\Sigma = \text{var}(\mathbf{X})$. Now $\text{var}(\mathbf{Y}) = I$, so the variables are uncorrelated. Hence there are no longer any linear regression relationships between variables. The $\mathbf{Y}$ variables also have no principal component information. Based on variable $\mathbf{Y}$ we then create a Fisher information matrix for $\mathbf{Y}$, denoted $J_{\mathbf{Y}}$.

In the second stage of "whitening," we create an orthogonal matrix Γ using the eigenanalysis of $J_{\mathbf{Y}}$. The new vector $\mathbf{Z} = \Gamma^{\mathrm{T}}\mathbf{Y} = \Gamma^{\mathrm{T}}\Sigma^{-1/2}\mathbf{X}$ has a diagonal Fisher information $J_{\mathbf{Z}}$. The diagonal entries of this matrix measure the information in each $\mathbf{Z}$ coordinate. We then use the $\mathbf{Z}$ coordinates with the greatest information in the data analysis.

To be more specific, let $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ be a d-dimensional random vector. Let density f have mean zero. We create the location family density $f(\mathbf{x} - \boldsymbol{\theta})$ with mean $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^{\mathrm{T}}$ and covariance matrix $\Sigma_{\mathbf{X}}$. In this article, we assume the density $f(\cdot)$ has continuous second derivative and that it satisfies the following regularity conditions that are standard in likelihood analysis:

$$\int f(\mathbf{x} - \boldsymbol{\theta})\,\mathrm{d}\mathbf{x} = 1, \quad \int \frac{\partial f(\mathbf{x} - \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\,\mathrm{d}\mathbf{x} = 0, \quad \int \frac{\partial^2 f(\mathbf{x} - \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}}\,\mathrm{d}\mathbf{x} = 0 \tag{1}$$

for any $\boldsymbol{\theta}$. Further, the integration and derivatives are exchangeable. Throughout this discussion we will also assume that the density f is sufficiently regular for the standard likelihood calculations to be valid for various conditional and marginal terms. For example, if $\mathbf{x}$ has one dimension, then the normal distribution and t-distribution satisfy the above regularity conditions. In fact, in applications, we will estimate f based on kernel smoothing by the normal density (see Section 4 for more details), which means our implementation always satisfies the necessary smoothness conditions.

Let

$$\nabla_{\boldsymbol{\theta}} f(\mathbf{x} - \boldsymbol{\theta}) = \left(\frac{\partial}{\partial \theta_1} f(\mathbf{x} - \boldsymbol{\theta}), \ldots, \frac{\partial}{\partial \theta_d} f(\mathbf{x} - \boldsymbol{\theta})\right)^{\mathrm{T}} = -\nabla_{\mathbf{x}} f(\mathbf{x} - \boldsymbol{\theta}).$$

The *(Fisher) information matrix* for density $f(\mathbf{x} - \theta)$ is then defined to be the variance matrix for the score vector

$$J_{\mathbf{X}} = \mathrm{E}\left[\nabla_{\boldsymbol{\theta}} \log f(\mathbf{X} - \boldsymbol{\theta}) \cdot \left(\nabla_{\boldsymbol{\theta}} \log f(\mathbf{X} - \boldsymbol{\theta})\right)^{\mathrm{T}}\right] = -\mathrm{E}\left[\frac{\partial^2 \log f(\mathbf{X} - \boldsymbol{\theta}))}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^{\mathrm{T}}}\right]. \tag{2}$$

Note that for the location family density $f(\mathbf{x} - \boldsymbol{\theta})$, the information matrix $J_{\mathbf{X}}$ does not depend on $\boldsymbol{\theta}$, that is, $\mathbf{X}$ and $\mathbf{X} + \boldsymbol{\theta}$ have the same information for any $\boldsymbol{\theta}$. Therefore, from now on, we will omit $\boldsymbol{\theta}$ in the density $f(\cdot)$.

We define the *(Fisher) score vector* for the density f to be $\nabla_{\mathbf{x}} \log f(\mathbf{x})$. Then the Fisher information matrix defined in (2) for density $f(\mathbf{x})$ can be also written as

$$J_{\mathbf{X}} = \mathrm{E}\left[\nabla_{\mathbf{X}} \log f(\mathbf{X}) \cdot \left(\nabla_{\mathbf{X}} \log f(\mathbf{X})\right)^{\mathrm{T}}\right] = \int \frac{\nabla_{\mathbf{x}} f \cdot \nabla_{\mathbf{x}} f^{\mathrm{T}}}{f(\mathbf{x})}\, \mathrm{d}\mathbf{x} = -\mathrm{E}\left[\frac{\partial^2 \log f(\mathbf{X}))}{\partial \mathbf{X} \partial \mathbf{X}^{\mathrm{T}}}\right]. \tag{3}$$

We will use the notation J_f instead of $J_{\mathbf{X}}$ when it is useful for clarity regarding the density involved.

An important feature of the Fisher score vector components is that they are simultaneously joint likelihood scores and conditional likelihood scores. For example, if ∇_j is the jth partial derivative, we can write

$$\nabla_{x_1} \log f(\mathbf{x}) = \frac{\nabla_{x_1} f(\mathbf{x})}{f(\mathbf{x})} = \frac{\nabla_{x_1} f(x_1|x_2, \ldots, x_d)}{f(x_1|x_2, \ldots, x_d)} = \nabla_{x_1} \log f(x_1|x_2, \ldots, x_d).$$

Later in this paper we will use this structure in various projection arguments.

The *Fisher information inequality* (Kagan et al., 1973) states that

$$J_{\mathbf{X}} \geq \Sigma_{\mathbf{X}}^{-1}, \tag{4}$$

and equality holds if and only if $f(\mathbf{x})$ is the multivariate normal density, where $A \geq B$ means that $A - B$ is a positive semi-definite matrix. Define the *standardized Fisher information matrix* for density $f(\mathbf{x})$ to be

$$W_{\mathbf{X}} = \Sigma_{\mathbf{X}}^{1/2} J_{\mathbf{X}} \Sigma_{\mathbf{X}}^{1/2}. \tag{5}$$

Hui & Lindsay (2010) called $W_{\mathbf{X}}$ (also denoted by W_f) the white noise matrix. Then we have the following white noise matrix inequality:

Proposition 2.1. *$W_f \geq I_d$ and equality occurs if and only if $f(\mathbf{x})$ is a multivariate normal density.*

This inequality is important for data analysis because of the transformation rules for Fisher information. We note that if $\mathbf{X}$ has density f, the Fisher information for any linear transformation $\mathbf{Y} = A\mathbf{X}$ having density g is

$$J_{\mathbf{Y}} = (A^{-1})^{\mathrm{T}} J_{\mathbf{X}} A^{-1}. \tag{6}$$

Using $A = \Sigma_{\mathbf{X}}^{-1/2}$ shows that the standardized vector $\mathbf{Y} = \Sigma_{\mathbf{X}}^{-1/2}\mathbf{X}$ satisfies $J_{\mathbf{Y}} = W_{\mathbf{X}}$, so the white noise matrix can be viewed as the information matrix for standardized data. Moreover, the white noise matrix for $\mathbf{Y}$ is

$$W_{\mathbf{Y}} = \Sigma_{\mathbf{Y}}^{1/2} J_{\mathbf{Y}} \Sigma_{\mathbf{Y}}^{1/2} = J_{\mathbf{Y}}, \tag{7}$$

since $\Sigma_{\mathbf{Y}} = I_d$. Hence $\mathbf{Y}$'s information has a lower bound of I based on Proposition 2.1. In addition, all the eigenvalues of $J_{\mathbf{Y}}$ are greater than or equal to 1.

Hui & Lindsay (2010) proposed using the white noise matrix $J_{\mathbf{Y}} = W_{\mathbf{X}}$ to detect the interesting low-dimensional features in the standardized data. Therefore, without loss of generality, we will assume that $\mathbf{Y} = \mathbf{X}$ has been standardized such that $\mathrm{E}(\mathbf{Y}) = 0$ and $\mathrm{cov}(\mathbf{Y}) = I_d$.

Proposition 2.2. *The ith diagonal term of $J_{\mathbf{Y}}$ for $\mathbf{Y}$ reaches the lower bound of 1 if and only if Y_i is marginally normally distributed and independent of any $\mathbf{Y}_{\langle i \rangle}$, with probability one over*

$\mathbf{Y}_{\langle i\rangle}$'s distribution, where $\mathbf{Y}_{\langle i\rangle} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_d)$. We will call such a Y_i a ***white noise coordinate****.*

In Section 3, we will provide an alternative proof and interpretation for the above result. We next use the white noise matrix to derive informative linear combinations of the $\mathbf{Y}$'s. Based on the eigendecomposition of $J_{\mathbf{Y}}$, we can write $\Gamma^{\mathrm{T}} J_{\mathbf{Y}} \Gamma = \Lambda$, where $\Gamma = (\gamma_1, \ldots, \gamma_d)$ is orthogonal and $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_d\}$; we assume $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Let $\mathbf{Z} = \Gamma^{\mathrm{T}}\mathbf{Y}$, with density $h(\mathbf{z})$. Then, based on (6), the Fisher information matrix for $\mathbf{Z}$ is $J_{\mathbf{Z}} = \Gamma^{\mathrm{T}} J_{\mathbf{Y}} \Gamma = \Lambda$.

Note that when some of the eigenvalues of $J_{\mathbf{Y}}$ are equal, the transformation Γ is not unique. In particular, if there are multiple eigenvalues of exactly 1, then there is a subspace of eigenvectors, any basis for which generates white noise variables. It is important that these variables are not just Gaussian, but also independent of the variables in the orthogonal subspace. This will be called a *white noise subspace*.

If $\lambda_{k+1} = 1$, then $\lambda_j = 1$, $j \geq k+1$. By Proposition 2.2, $Z_{k+1}, \ldots, Z_d$ are all white noise coordinates and thus do not carry any useful information. Therefore, we can simply use the "informative" projections $(Z_1, \ldots, Z_k)$ for further data analysis. In practice, we might also just use a smaller set of projections $(Z_1, \ldots, Z_m)$, where $m < k$ (such as $m = 1$ or 2), and we can then say they are the *m most informative projections*.

Note that in order to apply the above method to data, we will need to estimate Σ so as to find $\Sigma^{-1/2}$ and then estimate $J_{\mathbf{Y}}$ so as to estimate Γ. We will review the estimation methods given by Hui & Lindsay (2010) in Section 4.

Example 2.1 *Let $\mathbf{X} = (X_1, X_2, X_3)$, $X_1 \sim 0.5N(-3, 1) + 0.5N(3, 1)$ and X_2, X_3 be $N(0, 1)$. In addition, X_1, X_2, X_3 are independent. The covariance matrix for $\mathbf{X}$ is diagonal, so after standardization, $Y_2 = X_2$ and $Y_3 = X_3$. The information matrix for $\mathbf{Y}$ is diagonal, so $\Gamma = I$, and $\mathbf{Z} = \mathbf{Y}$. Therefore, $Z_2 = X_2$ and $Z_3 = X_3$ form a two-dimensional white noise subspace for this model and $Z_1 = aX_1$ is the interesting direction that shows the least Gaussian distribution, where $a = \{var(X_1)\}^{-1/2}$.*

3. SCORE ANOVA

In this section we show how the Fisher information matrix can be decomposed into a sum of positive definite covariance matrices, each of which corresponds to a separate "lack of fit" term. We will focus on decompositions that provide information about symmetries of densities and the independence of coordinates, as well as the Markov dependence between variables.

We will let $U(x) = \nabla_x \log f(x)$ and we will say that $U(x) = U_1(x) + U_2(x)$ is an *orthogonal decomposition* of U if $\mathrm{E}(U_1) = \mathrm{E}(U_2) = 0$ and if $\mathrm{E}(U_1 U_2') = 0$. A consequence of having an orthogonal decomposition is that if we let $J_1 = \mathrm{E}(U_1 U_1')$ and $J_2 = \mathrm{E}(U_2 U_2')$ be the information matrices corresponding to U_1 and U_2, then $J_f = J_1 + J_2$ is an information decomposition.

3.1. The Normality Decomposition

Our first decomposition, the *normality decomposition*, generates as a corollary the original information inequality. We assume that we have a standardized $\mathbf{Y}$ variable ($\mathrm{E}(\mathbf{Y}) = 0$ and $\mathrm{var}(\mathbf{Y}) = I$). Let $\phi(\mathbf{y})$ denote the standard normal density.

Proposition 3.1. *The scores $U_1 = \nabla \log(\phi(\mathbf{y}))$ and $U_2 = U - U_1$ are an orthogonal decomposition of U. The Fisher information for $\mathbf{Y}$ can therefore be decomposed as $J_f = J_{n\mathrm{lof}} + I$, where $J_{n\mathrm{lof}}$, the normality lack of fit matrix, is*

$$J_{n\mathrm{lof}} = E(U_2 U_2^{\mathrm{T}}) = E\left[\nabla \log(f(\mathbf{Y})) - \nabla \log(\phi(\mathbf{Y}))\right]\left[\nabla \log(f(\mathbf{Y})) - \nabla \log(\phi(\mathbf{Y}))\right]^{\mathrm{T}}.$$

We have $J_{n\mathrm{lof}} = 0$ if and only if $U_2 = 0$ a.e.; that is, if and only if f is standard normal.

Proof. Note that $\nabla \log(\phi(\mathbf{y})) = -\mathbf{y}$. Then the cross product terms in the expansion of $J_{n\mathrm{lof}}$ are

$$\begin{aligned}
\mathrm{E}\left[\nabla \log f(\mathbf{Y})\nabla^{\mathrm{T}} \log(\phi(\mathbf{Y}))\right] &= \int \nabla f(\mathbf{y})\nabla^{\mathrm{T}} \log(\phi(\mathbf{y}))\,\mathrm{d}\mathbf{y} \\
&= \int \nabla_{\boldsymbol{\theta}} f(\mathbf{y}-\boldsymbol{\theta})\mathbf{y}^{\mathrm{T}}\,\mathrm{d}\mathbf{y}\big|_{\boldsymbol{\theta}=0} \\
&= \nabla_{\boldsymbol{\theta}} \int f(\mathbf{y}-\boldsymbol{\theta})\mathbf{y}^{\mathrm{T}}\,\mathrm{d}\mathbf{y}\big|_{\boldsymbol{\theta}=0} \\
&= \nabla_{\boldsymbol{\theta}}\boldsymbol{\theta}^{\mathrm{T}}\big|_{\boldsymbol{\theta}=0} = I_d.
\end{aligned}$$

Therefore,

$$\mathrm{E}(U_2U_1^{\mathrm{T}}) = \mathrm{E}\left[\nabla \log f(\mathbf{Y})\nabla^{\mathrm{T}} \log(\phi(\mathbf{Y}))\right] - \mathrm{E}\left[\nabla \log(\phi(\mathbf{Y}))\nabla^{\mathrm{T}} \log(\phi(\mathbf{Y}))\right] = I_d - I_d = 0.$$

■

Proposition 2.1 is a corollary to Proposition 3.1 . We also note that the eigenvectors of $J_{\mathbf{Y}}$ are identical to those of $J_{n\mathrm{lof}}$, and so generate the same analysis as in Hui & Lindsay (2010).

Remark 3.1. *It is beyond our subject here but the form of $J_{n\mathrm{lof}}$ suggests a number of possible generalizations of white noise analysis. One could replace the density ϕ with g, where g corresponded to a non-Gaussian density. (Orthogonality may not hold, but it still measures lack of fit.) Or one could compare two samples by letting g be the density for the second sample.*

3.2. The Conditional/Marginal Decomposition

Our second information decomposition shows that the information matrix can be used to diagnose whether variables, or linear combinations of variables, are independent. We prove results for the original $\mathbf{X}$ variables, but they apply as well to the transformed $\mathbf{Y}$ and $\mathbf{Z}$ variables. Let $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{X}_1$ is r dimensional, $\mathbf{X}_2$ is s dimensional, and $r + s = d$. Let

$$J_{\mathbf{X}} = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix},$$

where J_{11} is the r by r matrix corresponding to $\mathbf{X}_1$.

Our second decomposition of U is based on the standard marginal-conditional factorization of the density:

$$U = \nabla \log f(\mathbf{x}_1, \mathbf{x}_2) = \nabla \log f(\mathbf{x}_2|\mathbf{x}_1) + \nabla \log f(\mathbf{x}_1).$$

If we let $U_1 = \nabla \log f(\mathbf{x}_2|\mathbf{x}_1)$ and $U_2 = \nabla \log f(\mathbf{x}_1)$, we can show

$$\mathrm{E}\left[\nabla \log f(\mathbf{X}_2|\mathbf{X}_1)\nabla^{\mathrm{T}} \log(f(\mathbf{X}_1))\right] = 0.$$

This is easily proved by carrying out the expectation conditionally on $\mathbf{X}_1$, and using the zero mean property of scores.

Therefore we have an orthogonal decomposition $U = U_1 + U_2$:

$$\begin{pmatrix} \nabla_1 \log f(\mathbf{x}) \\ \nabla_2 \log f(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} \nabla_1 \log f(\mathbf{x}_2|\mathbf{x}_1) \\ \nabla_2 \log f(\mathbf{x}_2|\mathbf{x}_1) \end{pmatrix} + \begin{pmatrix} \nabla_1 \log f(\mathbf{x}_1) \\ 0 \end{pmatrix}.$$

As a consequence, we can decompose the information matrix for $\mathbf{X}$ in the form

$$J_{\mathbf{X}} = J_{i\mathrm{lof}} + \begin{pmatrix} J_{\mathbf{X}_1} & 0 \\ 0 & 0 \end{pmatrix},$$

where $J_{\mathbf{X}_1}$ is the marginal information in $\mathbf{X}_1$. The *independence lack of fit* matrix

$$J_{i\mathrm{lof}} = \mathrm{var}[\nabla \log f(\mathbf{X}_2|\mathbf{X}_1)] = \begin{pmatrix} (J_{i\mathrm{lof}})_{11} & (J_{i\mathrm{lof}})_{12} \\ (J_{i\mathrm{lof}})_{21} & (J_{i\mathrm{lof}})_{22} \end{pmatrix} \tag{8}$$

has $(J_{i\mathrm{lof}})_{11} = 0$ if and only if $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent. Note that $(J_{i\mathrm{lof}})_{11} = J_{11} - J_{\mathbf{X}_1}$, and the right hand side is readily estimated from the data using the techniques of Hui & Lindsay (2010).

Similarly, based on the equation $\nabla \log f(\mathbf{x}_1, \mathbf{x}_2) = \nabla \log f(\mathbf{x}_1|\mathbf{x}_2) + \nabla \log f(\mathbf{x}_2)$, we obtain the decomposition

$$J_f = \mathrm{var}\big[\nabla \log f(\mathbf{X}_1|\mathbf{X}_2)\big] + \begin{pmatrix} 0 & 0 \\ 0 & J_{\mathbf{X}_2} \end{pmatrix}.$$

The above two decompositions can be used to prove the following proposition (Carlen, 1989; Kagan & Landsman, 1997):

Proposition 3.2.

$$J_{11} \geq J_{\mathbf{X}_1}, \quad J_{22} \geq J_{\mathbf{X}_2} \tag{9}$$

where the equalities both hold if and only if $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are independent. In addition, in (9), if one of the equalities holds, the other must hold, too.*

More generally, if $J_{\mathbf{X}}$ is a block diagonal matrix, with blocks corresponding to subvectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ then the subvectors are candidates to be independent of each other, and one can test if any one subvector is independent of the rest by comparing its block of $J_{\mathbf{X}}$ with the corresponding marginal information. In particular, if $J_{\mathbf{X}}$ is diagonal, then the variables are candidates to be mutually independent.

This proposition has been called the "superadditivity" of Fisher information. Hui & Lindsay (2010) added a conditional interpretation by pointing out that $J_{\mathbf{X}}(i,i)$ is a weighted average of $J_{X_i|\mathbf{X}_{\langle i\rangle}}$, the Fisher information values for the conditional distribution $f(x_i \mid \mathbf{x}_{\langle i\rangle})$, where $\mathbf{x}_{\langle i\rangle} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$.

Based on Proposition 3.2, we can also see that in fact $J_{\mathbf{X}}(i,i) \geq J_{X_i}$, where $J_{\mathbf{X}}(i,i)$ is the (i,i) element of $J_{\mathbf{X}}$.

Using the result of Proposition 3.2, we can also provide a proof and explanation different from that in Hui & Lindsay (2010) for Proposition 2.2. Assuming $\mathbf{Y}$ is standardized, based on Proposition 2.1 for $d = 1$, we know that $J_{Y_i} \geq 1$ with equality if and only if Y_i has a normal density. If the ith diagonal term of $J_{\mathbf{Y}}$ reaches the lower bound 1, then $1 \leq J_{Y_i} \leq J_{\mathbf{Y}}(i,i) = 1$. Therefore, $J_{Y_i} = 1$ and Y_i has a marginal normal density. In addition, since $J_{Y_i} = J_{\mathbf{Y}}(i,i)$, Y_i is independent of $\mathbf{Y}_{\langle i\rangle}$ based on Proposition 3.2.

3.3. Observed Information and Markov Independence

The observed Fisher information matrix provides a second stringent test of independence. Recall the information identity

$$J_{\mathbf{X}} = \mathrm{E}\big[\nabla_{\mathbf{X}} \log f \cdot (\nabla_{\mathbf{X}} \log f)^{\mathrm{T}}\big] = \mathrm{E}\big[-\nabla^2 \log f(\mathbf{X})\big]. \tag{10}$$

The negative Hessian in the last expectation is often called the observed Fisher information. It turns out to be diagnostic for the Markov independence of variables. Recall that a pair of variables $\mathbf{X}_1, \mathbf{X}_2$ are said to be *conditionally independent* given $\mathbf{X}_3$ if

$$\log(f(\mathbf{x}_1, \mathbf{x}_2|\mathbf{x}_3)) = \log(f(\mathbf{x}_1|\mathbf{x}_3)) + \log f(\mathbf{x}_2|\mathbf{x}_3).$$

Proposition 3.3. *Suppose* $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}}, \mathbf{X}_3^{\mathrm{T}})^{\mathrm{T}}$. *Then*

$$\frac{\partial^2 \log f(\mathbf{x})}{\partial \mathbf{x}_1 \partial \mathbf{x}_2^{\mathrm{T}}} = 0, \text{ a.e.,}$$

if and only if $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are conditionally independent given* $\mathbf{X}_3$.

Proof. If $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent given $\mathbf{X}_3$, then

$$f(\mathbf{x}) = f((\mathbf{x}_1, \mathbf{x}_2) \mid \mathbf{x}_3) f(\mathbf{x}_3) = f(\mathbf{x}_1 \mid \mathbf{x}_3) f(\mathbf{x}_2 \mid \mathbf{x}_3) f(\mathbf{x}_3).$$

Therefore,

$$\log f(\mathbf{x}) = \log\{f(\mathbf{x}_1 \mid \mathbf{x}_3)\} + \log\{f(\mathbf{x}_2 \mid \mathbf{x}_3)\} + \log\{f(\mathbf{x}_3)\}.$$

Then, we have

$$\frac{\partial^2 \log f(\mathbf{x})}{\partial \mathbf{x}_1 \partial \mathbf{x}_2^{\mathrm{T}}} = 0.$$

In addition, if $\partial^2 \log f(\mathbf{x})/(\partial \mathbf{x}_1 \partial \mathbf{x}_2^{\mathrm{T}}) = 0$, then

$$\frac{\partial^2 \log f(\mathbf{x})}{\partial \mathbf{x}_1 \partial \mathbf{x}_2^{\mathrm{T}}} = \frac{\partial^2 \log f((\mathbf{x}_1, \mathbf{x}_2) \mid \mathbf{x}_3)}{\partial \mathbf{x}_1 \partial \mathbf{x}_2^{\mathrm{T}}} = 0.$$

Then, $\log f((\mathbf{x}_1, \mathbf{x}_2) \mid \mathbf{x}_3) = a(\mathbf{x}_1, \mathbf{x}_3) + b(\mathbf{x}_2, \mathbf{x}_3)$ for some functions $a(\cdot)$ and $b(\cdot)$ and thus $f((\mathbf{x}_1, \mathbf{x}_2) \mid \mathbf{x}_3) = \exp\{a(\mathbf{x}_1, \mathbf{x}_3)\} \exp\{b(\mathbf{x}_2, \mathbf{x}_3)\}$. Therefore, $\mathbf{X}_1$ and $\mathbf{X}_2$ are independent given $\mathbf{X}_3$. ■

Thus structural zeroes of the observed information matrix are diagnostic for the conditional independency relationships of the variables. In particular, if there are no variables in $\mathbf{X}_3$, then having structural zeroes in the (1, 2) block proves the independence of $\mathbf{X}_1$ and $\mathbf{X}_2$.

3.4. Relationship to Independent Components

The results of the preceding subsection can be used to show that the Fisher information matrix provides meaningful information about independent component analysis (Jutten & Hérault, 1991; Common, 1994). We note that Common (1994) has had over 6,000 citations as of this writing.

Definition 3.1. *We say* $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ *is generated by an independent component analysis (ICA) model (Jutten & Hérault, 1991; Common, 1994) if*

$$\mathbf{X} = A\mathbf{Z} = \sum_{j=1}^{d} \mathbf{a}_j Z_j, \tag{11}$$

where $A = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ is a $d \times d$ nonsingular matrix, $\mathbf{Z} = (Z_1, \ldots, Z_d)^{\mathrm{T}}$, and $Z_1, \ldots, Z_d$ are mutually independent with unit variance.

In the ICA model, $\mathbf{X}$ is a vector of observed data, $\mathbf{Z}$ is a vector of the latent independent components, and A is the unknown mixing matrix. Here, for simplicity, we assume that A is a square nonsingular matrix, but this assumption can be relaxed in some situations.

The ICA analysis is closely related to the *cocktail-party problem* and related methods are sometimes called *blind source separation* or *blind signal separation*. The "source" means an original signal, that is, the independent component, such as the speaker in a cocktail party problem. We cannot directly record the original signal $\mathbf{Z}$. Instead, we only record the mixed signal $\mathbf{X}$, which is a linear mixture of original signals. The goal of the ICA model is to recover the original signals (such as the original speaker signals in a cocktail-party problem). Therefore, the ICA model tries to estimate both A and $\mathbf{Z}$ from an observed $\mathbf{X}$ under some conditions, that is, find a demixing matrix W such that $W\mathbf{X}$ is an estimate of $\mathbf{Z}$. See Hyvärinen & Oja (2000) for an introduction to the ICA model and its applications.

It is important to recognize that there are identifiability issues with the ICA model. The matrix A can be identified, up to the permutation, but only when at most one of the sources $\mathbf{Z}$ is Gaussian. That is, multiple Gaussian sources are intrinsically unseparable, and so an information analysis provides direct goodness-of-fit information about identifiability through its analysis of normality.

Based on Proposition 3.3 we have the following result.

Proposition 3.4. *For a random vector $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$, if $X_1, \ldots, X_d$ are mutually independent, then $J_{\mathbf{X}}$ is a diagonal matrix.*

Therefore, if $\mathbf{Y} = \Sigma^{-1/2}\mathbf{X}$ has a diagonal information matrix $J_{\mathbf{Y}}$, it is possible that $(Y_1, \ldots, Y_d)$ are mutually independent. If $J_{\mathbf{Y}}$ is not diagonal, we can transform the data to $\mathbf{Z} = \Gamma^{\mathrm{T}}\mathbf{Y}$, where Γ is the matrix of eigenvectors of $J_{\mathbf{Y}}$. The new vector $\mathbf{Z}$ has a diagonal white noise matrix $J_{\mathbf{Z}}$ with the diagonal entries being the eigenvalues of $J_{\mathbf{Y}}$. The transformation $\mathbf{Z} = \Gamma^{\mathrm{T}}\Sigma^{-1/2}\mathbf{X}$ has therefore brought the variables through two stages of whitening; now the covariance matrices of both vector $\mathbf{Z}$ and the score vector U are diagonal, and the elements of $\mathbf{Z}$ are twofold candidates to be mutually independent.

Proposition 3.5. *If the data $\mathbf{X}$ is generated by an independent component analysis model with covariance matrix Σ, and the eigenvalues of its white noise matrix $W_{\mathbf{X}}$ are distinct, then the transformed variables $\mathbf{Z} = \Gamma^{\mathrm{T}}\Sigma^{-1/2}\mathbf{X}$ are the independent component variables, up to permutation, where Γ is the matrix of eigenvectors of $W_{\mathbf{X}}$.*

Proof. If the white noise eigenvalues are unique, then the $\mathbf{Z} = \Gamma^{\mathrm{T}}\Sigma^{-1/2}\mathbf{X}$ are the unique (up to permutation) variables that simultaneously diagonalize the covariance and the information matrix. Since the mixing matrix A also diagonalizes both, they must be equal. ■

We note that when the eigenvalues are distinct, white noise theory not only provides candidates for the independent components, it also provides, through matrix decomposition, methods to check whether independence actually holds for those candidates.

The problem becomes trickier when the eigenvalues are not distinct. Suppose that the white noise matrix eigenanalysis of $\mathbf{Y} = \Sigma^{-1/2}\mathbf{X}$ has K distinct eigenvalues λ_1 to λ_K, where λ_j has multiplicity m_j and a corresponding subspace H_j of eigenvectors. The above proposition states that if the eigenvalues are of multiplicity $m_j = 1$, and there exists an independent component analysis for $\mathbf{Y}$, then the matrix A has rows that correspond to the unique eigenvectors.

However, if A is identifiable, which implies that no more than one eigenvalue of $J_{\mathbf{Z}}$ is 1, and if some m_j is greater than one, then any linear combination of vectors in the subspace H_j is a candidate to be one of the rows of A. For example, if Z_1 and Z_2 were standardized independent t variables, an information analysis would not provide unique eigenvectors that would identify Z_1 and Z_2 as the independent coordinates. One could still use our decomposition theory to test whether the subspaces corresponding to different eigenvalues each generated independent subvectors, however.

If we consider the observed information matrix instead, we can establish a matrix method to identify the orthogonal transformation Γ whenever it is identifiable:

Proposition 3.6. *The variables* $\mathbf{Z}$ *formed under transformation* $\mathbf{Z} = \Gamma^{\mathrm{T}}\mathbf{Y} = \Gamma^{\mathrm{T}}\Sigma^{-1/2}\mathbf{X}$ *are independent components, where* Γ *is an orthonormal matrix, if and only if*

$$\nabla_{\mathbf{z}}^2 \log(f_{\mathbf{Z}}(\mathbf{z}))|_{\text{offdiag}} = \Gamma^{\mathrm{T}}\nabla_{\mathbf{y}}^2 \log f(\mathbf{y})\Gamma|_{\text{offdiag}} = 0, \text{ a.e.},$$

where B_{offdiag} *denotes all off-diagonal elements of matrix B. If independent components holds, and there is a unique demixing matrix A, then there is a unique* Γ *for which the displayed equation holds, and A is equivalent to* $\Gamma^{\mathrm{T}}\Sigma^{-1/2}$.

Proof. If the transformed variables $\mathbf{Z}$ are independent, then they are also Markov independent, and so the off-diagonals are zero by Proposition 3.5. Conversely, if the off-diagonals are zero then all pairs Z_i, Z_j are Markov independent. That is, the graph for the variables has no edges, and so the variables are independent. ■

If one lets $h(\Gamma)$ be the sum of the variances of the off-diagonal entries of $\Gamma^{\mathrm{T}}\nabla_{\mathbf{y}}^2 \log f(\mathbf{y})\Gamma$, then this proposition says that $\min_\Gamma h(\Gamma) = 0$ when the ICA model holds, and in this case, $\tilde{\Gamma} = \arg\min_\Gamma h(\Gamma)$ generates the independent component model via $A = \tilde{\Gamma}^{\mathrm{T}}\Sigma^{-1/2}$.

Of course if $\min_\Gamma h(\Gamma)$ is not zero, independent components is not satisfied. In this case rotational symmetry is an alternative hypothesis to the independent component model that can be used to explain the existence of non-Gaussian subspaces, with $m_j > 1$. We will examine this hypothesis in a later section.

3.5. The Markov Decomposition

The preceding independence decomposition shows that if the Fisher information matrix $J_{\mathbf{X}}$ is block diagonal then the corresponding sets of variables are candidates to be mutually independent. Further, we can test for this by examining the matrix $J_{i\text{lof}}$ of (8) for each block.

We next consider the interpretation of other blocks of zeroes in the off-diagonal regions of the matrix. That is, how a structure of the following form might arise:

$$J_{\mathbf{X}} = \begin{bmatrix} J_{11} & J_{12} & 0 \\ J_{21} & J_{22} & J_{23} \\ 0 & J_{32} & J_{33} \end{bmatrix}. \tag{12}$$

where $J_{\mathbf{X}}$ is the suitably partitioned Fisher information for $\mathbf{X} = (\mathbf{X}_1^{\mathrm{T}}, \mathbf{X}_2^{\mathrm{T}}, \mathbf{X}_3^{\mathrm{T}})^{\mathrm{T}}$. The observed Fisher information result in Proposition 3.3 indicates that a structural off-diagonal zero in the (1, 3) entries of observed information tells us that $\mathbf{X}_1$ and $\mathbf{X}_3$ are independent in their conditional relationship given $\mathbf{X}_2$. Hence an off-diagonal zero in $J_{\mathbf{X}}$ suggests this might be true. We now show how to construct an information decomposition that definitively tests this hypothesis, as well as providing a numerical measure of the degree of dependence when it exists.

Note that J_{13} is equal to the (1, 3) block of J_X. Therefore, we have the following result.

Proposition 3.7. *For the Fisher information partition of (12), we have the following results.*

a. If $\mathbf{X}_1$ *and* $\mathbf{X}_3$ *are conditionally independent given* $\mathbf{X}_2$*, then* $J_{13} = J_{31}^{\mathrm{T}} = 0$*;*

b. We have

$$J_{11} = J_{\mathbf{X}_1,\mathbf{X}_2}(1,1) + var\big[\nabla_1 \log(f(\mathbf{X}_3|\mathbf{X}_2,\mathbf{X}_1))\big],$$

where $J_{\mathbf{X}_1,\mathbf{X}_2}(1,\ 1) = var([\nabla_1 \log(f(\mathbf{X}_1,\mathbf{X}_2))]$ *is the* $(1,\ 1)$ *block matrix of* $J_{\mathbf{X}_1,\mathbf{X}_2}$ *corresponding to* $\mathbf{X}_1$ *and the equality* $J_{11} = J_{\mathbf{X}_1,\mathbf{X}_2}(1,1)$ *(i.e.,* $var[\nabla_1 \log(f(\mathbf{X}_3|\mathbf{X}_2,\mathbf{X}_1))] = 0$*) occurs if and only if* $\mathbf{X}_1$ *and* $\mathbf{X}_3$ *are conditionally independent given* $\mathbf{X}_2$*;*

c. We have

$$J_{33} = J_{\mathbf{X}_2,\mathbf{X}_3}(2,2) + var\big[\nabla_3 \log(f(\mathbf{X}_1|\mathbf{X}_2,\mathbf{X}_3))\big],$$

where $J_{\mathbf{X}_2,\mathbf{X}_3}(2,\ 2) = var([\nabla_3 \log(f(\mathbf{X}_2,\mathbf{X}_3))]$ *is the* $(2,\ 2)$ *block matrix of* $J_{\mathbf{X}_2,\mathbf{X}_3}$ *corresponding to* $\mathbf{X}_3$ *and the equality* $J_{33} = J_{\mathbf{X}_2,\mathbf{X}_3}(2,2)$ *(i.e.,* $var[\nabla_3 \log(f(\mathbf{X}_1|\mathbf{X}_2,\mathbf{X}_3))] = 0$*) occurs if and only if* $\mathbf{X}_1$ *and* $\mathbf{X}_3$ *are conditionally independent given* $\mathbf{X}_2$*.*

Proof.

(a) The result is a corollary of Proposition 3.3.

(b) Note that $\log f(\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3) = \log f(\mathbf{x}_1,\mathbf{x}_2) + \log f(\mathbf{x}_3 \mid \mathbf{x}_1,\mathbf{x}_2)$ creates an orthogonal decomposition. Then, we have

$$J_{11} = J_{\mathbf{X}_1,\mathbf{X}_2}(1,1) + \mathrm{var}[\nabla_1 \log(f(\mathbf{X}_3|\mathbf{X}_2,\mathbf{X}_1))].$$

If $\mathrm{var}[\nabla_1 \log(f(\mathbf{X}_3|\mathbf{X}_2,\mathbf{X}_1))] = 0$, then $\nabla_1 \log(f(\mathbf{X}_3|\mathbf{X}_2,\mathbf{X}_1)) = 0$, a.e. (since it is a score function and has mean 0), that is, the conditional density of $\mathbf{X}_3 \mid \mathbf{X}_1,\mathbf{X}_2$ does not depend on $\mathbf{X}_1$. This in turn is equivalent to saying that $\mathbf{X}_1$ and $\mathbf{X}_3$ are conditionally independent given $\mathbf{X}_2$. The reverse result is obvious.

(c) This part can be proved similarly to part b).

■

3.6. Markov Dependency

A graphical model is a probabilistic model in which a graph is used to describe the conditional independence structure between random variables. In this section we describe how the Fisher Information matrix provides tools for graphical models. We first give an introduction to these concepts (Jordan, 1999; Jordan et al., 1999).

Graphical models play an important role in probability theory, Bayesian statistics, and the design and analysis of machine learning algorithms. There are two main branches of graphical models: Bayesian networks and Markov networks. Here we are interested in Markov networks.

Definition 3.2. *We will say that* X_s *and* X_t *are Markov independent if* X_s *and* X_t *are conditionally independent given* $\mathbf{X}_{\langle s,t\rangle}$*. We will say that a matrix* G *with* $(i,\ j)$*th element* $G_{i,j}$ *is a Markov dependency matrix if* $G_{st} = 0$ *whenever* X_s *and* X_t *are Markov independent.*

The inverse of the covariance matrix, Σ^{-1}, also called the precision matrix, is a Markov dependency matrix. Under the restrictive condition of multivariate normality, the off-diagonal zeroes are also conclusive about Markov dependency. We have now shown that $J_{\mathbf{X}}$ is also a Markov dependency matrix in Proposition 3.7, and the matrix decomposition we found can be used to provide a conclusive test of Markov dependency, irrespective of normality.

Proposition 3.8 *The Fisher information matrix* $J_{\mathbf{X}}$ *is a Markov dependency matrix, that is, if* X_s *and* X_t *are independent given* $\mathbf{X}_{\langle s,t\rangle}$*, then* $J_{s,t} = 0$*, where* $s \neq t$*.*

In a Markov dependency matrix, we know that every non-zero entry in J implies Markov dependence, while every zero entry is a candidate for Markov independence. One can create a more precise test for Markov independence based on the variance of $\nabla_1 \log(f(\mathbf{x}_3|\mathbf{x}_2, \mathbf{x}_1))$ or $\nabla_3 \log(f(\mathbf{x}_1|\mathbf{x}_2, \mathbf{x}_3))$ as seen in Proposition 3.7(a) and (b). One could also use the variance of the off-diagonal entries of the observed information matrix.

3.7. Measuring Lack of Spherical Symmetry

In this section, we return to the issue of repeated eigenvalues for $J_{\mathbf{Y}}$. We have already seen that the observed information matrix provides information about the independent components model in this case. We now show that there exists a decomposition of the information matrix that provides conclusive evidence about the alternative hypothesis of rotational symmetry in this situation.

Definition 3.3 *A $(d \times 1)$ random vector $\mathbf{X}$ is said to be spherically symmetric if the distribution of $\mathbf{X}$ is identical to $Q\mathbf{X}$ for all $(d \times d)$ orthogonal matrices Q.*

Maxwell (1860), Bartlett (1934), and Hartman & Wintner (1940) are the three earliest papers pertaining to spherically symmetric distributions. Spherically symmetric distributions are appearing with increasing frequency in the literature and are the usual assumptions used in many dimension reduction algorithms (Li, 1991; Cook & Nachtsheim, 1994; Cook & Li, 2002).

Proposition 3.9. *If $\mathbf{Y}$ is spherically symmetric with covariance Σ and white noise matrix $J_{\mathbf{Y}}$, then both Σ and $J_{\mathbf{Y}}$ are multiples of the identity matrix.*

Proof. The invariance property of the spherically symmetric distribution implies that $Q'\Sigma Q = \Sigma$ for any orthogonal Q. If we let Q be the eigenvectors of Σ, then we see that Σ is diagonal. Clearly the symmetry means that all the diagonal elements are equal. A similar argument, using the transformation properties of the Fisher information in (6), shows that it is diagonal with equal entries on the diagonal. Hence the white noise matrix must also be a multiple of the identity matrix. ■

From this result, it is clear that we cannot distinguish between spherically symmetric densities and i.i.d. coordinate densities based only on the matrices Σ and $J_{\mathbf{Y}}$. However, these two classes of densities have very little overlap. Maxwell (1860), Bartlett (1934), and Hartman & Wintner (1940) proved that the spherically symmetric random vector $\mathbf{X}$ has independent components if and only if $\mathbf{X}$ follows a multivariate normal distribution.

There is conclusive evidence that the distribution is not spherically symmetric if the eigenvalues of the Fisher information are not all equal. Thus we consider how we might identify a radial symmetry model when the eigenvalues are equal.

Definition 3.4 *A $(d \times 1)$ random vector $\mathbf{X}$ with mean $\boldsymbol{\mu}$ and covariance Σ is said to be elliptically symmetric (Kelker, 1970) if its density function f has the structure*

$$f(\mathbf{x}) = k \cdot g\big((\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\big),$$

where k is a constant, $\boldsymbol{\mu}$ is the mean of $\mathbf{X}$, and Σ is the covariance matrix of $\mathbf{X}$.

Note that if $\boldsymbol{\mu} = 0$ and Σ is an identity matrix, then $\mathbf{X}$ has spherical symmetry. If $\mathbf{X}$ is elliptically symmetric with covariance matrix Σ and mean $\boldsymbol{\mu}$, then $\mathbf{Y} = \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ is spherically symmetric. In addition, based on Proposition 3.9, the white noise matrix, $J_{\mathbf{Y}}$, is diagonal and has equal eigenvalues. If the diagonal entries of the matrix $J_{\mathbf{Y}}$ equal one, then we know that $\mathbf{Y}$ is normal, and the elements are independent variables. However, if the diagonal entries are not one, then the variables are not normal.

Therefore, if we carry out a white noise analysis, and the eigenvalues are distinct, the uniquely defined $\mathbf{Z}$ variables are candidates for independent components, but not for spherically symmetric components since a spherically symmetric density has equal eigenvalues. If some of the eigenvalues are the same, then the corresponding subspace is a candidate either for being spherically symmetric or for having independent components, or neither. We can test for the presence of spherical symmetry by using the following device.

We will assume the variables are *quasi-independent*, so that the white noise matrix and the information matrix are equidiagonal (diagonal with equal diagonal values), and we have non-unique eigenvectors. Given a non-zero vector $\mathbf{x}$, let $r^2 = \mathbf{x}^{\mathrm{T}}\mathbf{x}$, and let $P_{\mathbf{x}}$ denote the projection matrix onto vector $\mathbf{x}$, namely

$$P_{\mathbf{x}} = \mathbf{x}(r^2)^{-1}\mathbf{x}^{\mathrm{T}} = r^{-2}\mathbf{x}\mathbf{x}^{\mathrm{T}}.$$

We call $U_1 = P_{\mathbf{x}}\nabla \log f(\mathbf{x})$ the *radially projected score function*, and let $U_2 = U - U_1$. We note that U_1 and U_2 are in fact orthogonal for every fixed $\mathbf{x}$, and so we have an orthogonal decomposition.

One characterization of a spherically symmetric density is that $\nabla \log f(\mathbf{x}) = c(\mathbf{x})\mathbf{x}$, where $c(\mathbf{x})$ is a scalar function. For example, for the i.i.d. normal, the log density gradient is $-\mathbf{x}$. It follows that $U_2 = \nabla \log f(\mathbf{x}) - P_{\mathbf{x}}\nabla \log f(\mathbf{x}) = 0$ a.e. is equivalent to the hypothesis that the density f is spherically symmetric.

Let us then define the *spherical symmetry lack of fit matrix* to be

$$J_{s\mathrm{lof}} = \mathrm{E}\left[\nabla \log f(\mathbf{X}) - P_{\mathbf{X}}\nabla \log f(\mathbf{X})\right]\left[\nabla \log f(\mathbf{X}) - P_{\mathbf{X}}\nabla \log f(\mathbf{X})\right]^{\mathrm{T}}.$$

Further, let $J_{\mathrm{sph}} = \mathrm{E}[P_{\mathbf{X}}\nabla \log f(\mathbf{X})][P_{\mathbf{X}}\nabla \log f(\mathbf{X})]^{\mathrm{T}}$ be the spherical information matrix.

Proposition 3.10. *We have*

$$J_{\mathbf{X}} = J_{\mathrm{sph}} + J_{s\mathrm{lof}}.$$

Further, $J_{s\mathrm{lof}} = 0$ if and only if the density f is spherically symmetric.

One can carry out a further decomposition in order to understand the information content of J_{sph}. We now let $U_1 = \nabla \log \phi(\mathbf{x})$ and $U_2 = P_{\mathbf{x}}\nabla \log f(\mathbf{x}) - U_1$. If f is spherically symmetric, then $P_{\mathbf{x}}\nabla \log f(\mathbf{x}) = \nabla \log f(\mathbf{x})$, and so this is an orthogonal decomposition from our first result. That is, it can be orthogonally decomposed into $J_{\mathrm{sph}} = J_{g\mathrm{lof}} + I$, where the Gaussian lack of fit matrix

$$J_{g\mathrm{lof}} = E\{P_{\mathbf{x}}\nabla \log f(\mathbf{x}) + \mathbf{x}\}\{P_{\mathbf{x}}\nabla \log f(\mathbf{x}) + \mathbf{x}\}^{\mathrm{T}}$$

is a matrix that is zero if the density is normal. If the data is spherically symmetric but not normal, then its white noise matrix is equidiagonal.

4. ESTIMATION OF WHITE NOISE MATRIX

4.1. Step 1: Kernel Smoothing

To perform inference based on the preceding theory, we will need to estimate the white noise matrix for an unknown density f. The solution of Hui & Lindsay (2010) was to use a kernel density estimator. They showed that by using the normal kernel one could obtain a consistency theory for the white noise subspaces that was valid even when the bandwidth was held fixed.

Given the well known challenges of kernel density estimation in higher dimensions this is an important result. We here extend their results to our new theory.

Given the standardized data $\mathbf{y}_1, \ldots, \mathbf{y}_n$ from density f, suppose we use the kernel density estimator $\hat{f}(\mathbf{y}^*) = \sum n^{-1} K_H(\mathbf{y}_i - \mathbf{y}^*)$ to replace f in our matrix calculations, where $K_H(t_1, \ldots, t_d) = k_h(t_1)k_h(t_2)\cdots k_h(t_d)$ is an independent kernel. For any fixed bandwidth H, this estimator is consistent for estimating

$$f^*(\mathbf{y}^*) = \int k_H(\mathbf{y} - \mathbf{y}^*)\,\mathrm{d}F(y) = \int k_H(\mathbf{y} - \mathbf{y}^*) f(\mathbf{y})\,\mathrm{d}\mathbf{y},$$

but not $f(\mathbf{y})$. If we wish to have bandwidth-free consistency of our method, we need to verify that the change from f to f^* does not change the key information properties we have described.

If $\mathbf{Y} = \Sigma^{-1/2}\mathbf{X}$ has density f, then the random variable that corresponds to f^* is $\mathbf{Y}^* = \mathbf{Y} + \varepsilon$, where ε is a vector of i.i.d. errors from the density $k_h(\cdot)$. Our first point is that it should be highly desirable to use a normal kernel k for smoothing, as a white noise coordinate in $\mathbf{Y}$ remains a white noise coordinate in $\mathbf{Y}^*$. This statement can be generalized as follows. If ε is distributed as a multivariate normal vector with mean zero and covariance matrix H, independent of $\mathbf{Y}$, then $\mathbf{Y}^* = \mathbf{Y} + \varepsilon$ has a white noise subspace identical to that of $\mathbf{Y}$. In particular, if $\mathbf{a}'\mathbf{Y}$ is normally distributed, so is $\mathbf{a}'\mathbf{Y}^*$, and so the white noise analysis is unchanged.

The effect of normal smoothing on the independent component analysis is a bit more subtle. Suppose $A\mathbf{X} = \Gamma^{\mathrm{T}}\Sigma^{-1/2}\mathbf{X} = \mathbf{Z}$ is an independent component representation. If $\mathbf{X}$ has been standardized to variable $\mathbf{Y} = \Sigma^{-1/2}\mathbf{X}$, and we apply the i.i.d. normal smoothing to $\mathbf{Y}$ to get $\mathbf{Y}^*$, then it is clear that $\mathbf{Z}^* = \Gamma^{\mathrm{T}}(\mathbf{Y} + \varepsilon) = \mathbf{Z} + \Gamma^{\mathrm{T}}\varepsilon$ also represents independent components. This is because $\Gamma^{\mathrm{T}}\varepsilon$ is again distributed as i.i.d. normal, and we are convolving two independent vectors.

However, we should note that if one were to apply i.i.d. normal smoothing to $\mathbf{X}$, instead of $\mathbf{Y}$, and used $\mathbf{X}^* = \mathbf{X} + \varepsilon$ for analysis, then this argument does not hold. This is because in the transformed variable $\mathbf{Z}^* = \Gamma\Sigma^{-1/2}(\mathbf{X} + \varepsilon)$, the vector $\Sigma^{-1/2}\varepsilon$ will only under special circumstances be independent normals. One special circumstance occurs when ε is multivariate normal with mean 0 and variance $h^2\Sigma$.

The effect of kernel smoothing on the spherical symmetry analysis is straightforward. Suppose that $\mathbf{Y}$ has a spherically symmetric distribution. Then for every orthogonal transformation Λ, the distribution of $\Lambda\mathbf{Y}$ equals that of $\mathbf{Y}$. It therefore follows that $\Lambda(\mathbf{Y} + \varepsilon) =_{\text{dist}} \mathbf{Y} + \varepsilon$, and so $\mathbf{Y}^*$ is also spherically symmetric.

Therefore, we have the following results.

Proposition 4.1. *Suppose independent normal kernel smoothing has been applied to the standardized variable* $\mathbf{Y}$*, giving the smoothed density* $f^*(y_1, \ldots, y_d)$*. Then*

(a) $Y_1, \ldots, Y_k$ *are mutually independent and independent of* $(Y_{k+1}, \ldots, Y_d)$*, where* $1 \le k \le d$*, if and only if* $Y_1^*, \ldots, Y_k^*$ *are mutually independent and independent of* $(Y_{k+1}^*, \ldots, Y_d^*)$*. This result extends to arbitrary blocks of variables.*

(b) $Z_1, \ldots, Z_k$ *are white noise subspace, that is,* $Z_1, \ldots, Z_k$ *are independent of* $Z_{k+1}, \ldots, Z_d$ *and have an independent multivariate normal distribution, if and only if* $Z_1^*, \ldots, Z_k^*$ *are a white noise subspace.*

(c) $\mathbf{Y}$ *is generated by an independent component analysis, if and only if* $\mathbf{Y}^*$ *is also generated by an independent component analysis. In addition, they have the same mixing matrix.*

(d) $\mathbf{Y}$ *is spherically symmetric if and only if* $\mathbf{Y}^*$ *is spherically symmetric.*

Based on the above proposition, we can see that the smoothed density preserves the white noise subspace, normality, independent components, and rotational symmetry properties, and so we should expect bandwidth free consistency results when we use normal kernel smoothing on the standardized variables.

Remark 4.1. *Unfortunately the Markov independence results are not preserved under kernel smoothing, and so any results regarding consistency of estimation would require that the bandwidth go to zero.*

4.2. Step 2: Density Square Transformation

If we estimate $J_{\mathbf{X}}$ in (3) by replacing the density f with a kernel density estimate $\hat{f}$, the integration will not have an explicit form. One could still apply a simulation method to estimate the information matrices involved. However, Hui & Lindsay (2010) proposed instead using a density square transformation. In addition to making the computation explicit, it offered a second advantage in that it robustified the method by downweighting outliers. In this section we address how the square transformation might affect the decomposition of information results.

We start by creating a new random variable $\mathbf{S}$ that has the density

$$f_{(2)}(\mathbf{s}) \equiv \frac{f^2(\mathbf{s})}{\int f^2(\mathbf{x})\,\mathrm{d}\mathbf{x}},$$

where $f(\mathbf{x})$ is the density of $\mathbf{X}$. We then estimate the information for the density $f_{(2)}(\mathbf{s})$, which we can denote as $J_{\mathbf{S}}$ or $J_{f_{(2)}}$.

As argued by Hui & Lindsay (2010), the square density $f_{(2)}(\mathbf{x})$ has the same contour lines as the original $f(\mathbf{x})$ and in particular the same peaks and valleys. The main changes are that the peaks become more accentuated in the square density, but the tails are downweighted. In addition, $\mathbf{X}$ is normal if and only if $\mathbf{S}$ is normal. Therefore, the density square transformation preserves the white noise subspace. Estimating the most informative directions for $\mathbf{S}$ therefore seems a useful surrogate for estimating the most informative directions for $\mathbf{X}$. What concerns us here is the effect of the transformation from $\mathbf{X}$ to $\mathbf{S}$ upon the dependency and symmetry properties we have described.

If we plug in $f_{(2)}$ into (5), then we have the inequality

$$J_{f_{(2)}} = \frac{\Sigma_{f_{(2)}}^{1/2} \int \nabla_{\mathbf{x}} f \cdot \nabla_{\mathbf{x}} f^{\mathrm{T}}\,\mathrm{d}\mathbf{x}\,\Sigma_{f_{(2)}}^{1/2}}{\int f^2(\mathbf{x})\,\mathrm{d}x} \geq \frac{1}{4} I_d, \tag{13}$$

where $\Sigma_{f_{(2)}}$ is the variance of $\mathbf{S}$. The equality holds if and only if the standardized $\mathbf{S}$ is multivariate normal.

Now $J_{\mathbf{S}}$ has an explicit form for estimation provided that we estimate f using

$$\hat{f}_H(\mathbf{x}) = \sum_{i=1}^{n} \frac{1}{n|H|} \phi_d\big(\mathbf{x} - \mathbf{x}_i, 0, H^2\big), \tag{14}$$

where $\phi_d(x, 0, H^2)$ is the d-dimensional multivariate normal density function with mean 0 and covariance H^2. See Hui & Lindsay (2010) for the explicit formula of the estimator of $J_{\mathbf{S}}$. Based on Proposition 2.9 of Zografos, Ferentinos, & Papaioannou (1989), we know that $J_{\hat{f}_H^2} \to J_{f_{(2)}}$ if H goes to zero (elementwise). In addition, Hui & Lindsay (2010) argued that $J_{\hat{f}_H^2}$ can be also considered as a direct measure of the non-normality of the kernel smoothed distribution

$$f_{(2)}^* = \frac{f^*(\mathbf{x})^2}{\int f^*(\mathbf{y})\,\mathrm{d}\mathbf{y}},$$

where

$$f^{*}(\mathbf{x}) = \int f(\mathbf{y}) \frac{1}{|H|} \phi_d(\mathbf{x} - \mathbf{y}, 0, H^2)\, \mathrm{d}\mathbf{y},$$

since $J_{\hat{f}^2_H}$ is a consistent estimator of $J_{f^*_{(2)}}$ *without* H going to zero.

We can now extend their results by noting that the squared density transformation preserves the properties we have been investigating in this paper.

Proposition 4.2 . *Suppose the random vector* $\mathbf{X} = (X_1, \ldots, X_d)^{\mathrm{T}}$ *has the density* $f(\mathbf{x})$ *and the random variable* $\mathbf{S} = (S_1, \ldots, S_d)^{\mathrm{T}}$ *has the square density* $f_{(2)}(\mathbf{s})$ *of* f. *Then we have the following results.*

(a) $X_1, \ldots, X_k$ *are mutually independent and independent of* $(X_{k+1}, \ldots, X_d)$, *where* $1 \le k \le d$, *if and only if* $S_1, \ldots, S_k$ *are mutually independent and independent of* $(S_{k+1}, \ldots, S_d)$.

(b) X_u *and* X_v *are independent given* $\mathbf{X}_{\langle u,v \rangle}$ *if and only if* S_u *and* S_v *are independent given* $\mathbf{S}_{\langle u,v \rangle}$.

(c) $\mathbf{X}$ *is generated by an independent component analysis if and only if* $\mathbf{S}$ *is also generated by an independent component analysis. In addition, they have the same mixing matrix.*

(d) $\mathbf{X}$ *is spherically symmetric if and only if* $\mathbf{S}$ *is spherically symmetric.*

Proof. The proofs for (a), (b), and (d) are trivial. Here, we just provide the proof for (c). If $\mathbf{X}$ is generated by an independent component analysis, then there exists a nonsingular matrix $A = (\mathbf{a}_1, \ldots, \mathbf{a}_d)^{\mathrm{T}}$, such that $A\mathbf{X} = \mathbf{Z}$, where the independent components $\mathbf{Z}$ has the density $f_1(z_1) f_2(z_2) \cdots f_d(z_d)$. Then the density of $\mathbf{X}$ is

$$f(\mathbf{x}) = f_1\left(\mathbf{a}_1^{\mathrm{T}}\mathbf{x}\right) f_2\left(\mathbf{a}_2^{\mathrm{T}}\mathbf{x}\right) \cdots f_d\left(\mathbf{a}_d^{\mathrm{T}}\mathbf{x}\right) |A|.$$

Therefore, the density of $\mathbf{S}$ is

$$f_{(2)}(\mathbf{s}) \propto f_1\left(\mathbf{a}_1^{\mathrm{T}}\mathbf{s}\right)^2 f_2\left(\mathbf{a}_2^{\mathrm{T}}\mathbf{s}\right)^2 \cdots f_d\left(\mathbf{a}_d^{\mathrm{T}}\mathbf{s}\right)^2.$$

Let $\mathbf{W} = A\mathbf{S}$, then the density of $\mathbf{W}$ is

$$g(\mathbf{w}) \propto f_1(w_1)^2 f_2(w_2)^2 \cdots f_d(w_d)^2.$$

Therefore, $W_1, \ldots, W_d$ are mutually independent and thus $\mathbf{S}$ is also generated by an independent component analysis with the same mixing matrix A^{-1} as $\mathbf{X}$. The converse can be proved similarly. ■

5. DISCUSSION

In this paper we have shown that the information content of the white noise matrix goes well beyond detection of Gaussianity or not. It can be used as a diagnostic for such features as Markov dependence, independent component structure, and spherical symmetry. We provided orthogonal decompositions of the white noise matrix that could be used to test the fit of these models. In particular, for the independent component analysis model, we proved that the white noise matrix can be an alternative tool to estimate the demixing matrix provided that the eigenvalues of the white noise matrix are distinct.

In this discussion we tackle the difficult question of how well the estimation methodologies can be adapted to data where the dimension d is large relative to the sample size n.

It is clear that some modifications need to be made in these cases. If there is insufficient data $\mathbf{X}$, and one uses the sample covariance $\hat{S}$ to estimate Σ, then the standardized data

$\mathbf{Y} = \hat{S}^{-1/2}\mathbf{X}$ (possibly using a generalized inverse) will have no interesting structure left for analysis. For example, if there are $n = d + 1$ data points $x_1, \ldots, x_{d+1}$ in d dimensional space, then the standardized data points $y_1, \ldots, y_{d+1}$ are exactly the corner points of a regular simplex. For example, when $d = 2$, the three $\mathbf{y}$ data points are on the corners of an equilateral triangle.

Hui & Lindsay (2010) did apply the white noise method to such sparse high-dimensional data by kernel smoothing the original $\mathbf{X}$ variables to get $\mathbf{X}^*$, then standardizing the $\mathbf{X}^*$ information matrix to form the white noise matrix. The covariance matrix for $\mathbf{X}^*$ is then $\Sigma + h^2 I$, which is invertible even when the sample covariance is used for Σ. As we noted above the Gaussian detection properties will be preserved over all h with this kind of smoothing, but the guaranteed independence detection properties will be lost. However, one might suppose that when $d > n$, there is little hope of detecting independence between variables anyway, as one cannot even estimate the full set of covariances. Hui & Lindsay (2010) did have some success in detecting mixtures of normal densities in their limited simulation of this smoothing methodology, but we think that application of white noise methodology in this domain requires more theoretical understanding and more simulation experience. It is clear that part of the understanding involves the interplay between principal components and white noise when the variables are not standardized.

As to future work, there is much to do to implement the theoretical goodness-of-fit decompositions described in this paper. In addition, there are more theoretical and methodological developments that are possible. Another possible step forward would be the analysis of "local" information to look for local structures of dependency and non-normality.

We have not discussed in this paper the possible applications of white noise analysis in a regression setting. However we have seen through this paper that the Fisher information matrix is directly useful for understanding the dependency relationships between variables. We believe that new dimension reduction techniques for covariate spaces will be a straightforward extension of the ideas in this paper.

ACKNOWLEDGEMENTS

The authors are grateful to the editors, especially Professor Changbao Wu, and the referees for their insightful comments on this paper. Bruce G. Lindsay's research was supported by NSF grant CCF 0936948.

BIBLIOGRAPHY

Bartlett, M. S. (1934). The vector representation of a sample. *Proceedings of the Cambridge Philosophical Society*, 30, 327–340.

Carlen, E. A. (1989). Superadditivity of Fisher's information and logarithmic Sobolev inequalities. *Journal of Functional Analysis*, 101, 194–211.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36, 287–314.

Cook, R. D. & Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics*, 30, 455–474.

Cook, R. D. & Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association*, 89, 592–599.

Friedman, J. H. & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C23, 881–889.

Hartman, P. & Wintner, A. (1940). On the spherical approach to the normal distribution law. *American Journal of Mathematics*, 62, 759–779.

Huber, P. (1985). Projection pursuit. *Annals of Statistics*, 13, 435–475.

Hui, G. & Lindsay, B. G. (2010). Projection pursuit via white noise matrices (with discussion). *Sankhya*, B72, 123–153.

Hyvärinen, A. & Oja, E. (2000). Independent component analysis: Algorithm and applications. *Neural Networks*, 13, 411–430.

Jones, M. & Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society, Series A*, 150, 1–36.

Jordan, M. I. (1999). *Learning in Graphical Models*. MIT Press, Cambridge, Massachusetts, London, England.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37, 183–233.

Jutten, C. & Hérault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24, 1–10.

Kagan, A. & Landsman, Z. (1997). Statistical meaning of Carlen's superadditivity of the Fisher information. *Statistics and Probability Letters*, 32, 175–179.

Kagan, A. M., Linnik, Y. V., & Rao, C. R. (1973). *Characterization Problems in Mathematical Statistics*. Wiley Series in Probability and Mathematical Statistics, No. 1. ISBN-10: 0471454214

Kelker, D. (1970). Distribution theory of spherical distributions and a location-scale parameter generalization. *Sankhya*, A32, 419–430.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86, 316–342.

Maxwell, J. C. (1860). Illustration of the dynamical theory of gases-Part I. On the motions and collisions of perfectly elastic bodies. *Taylor's Philosophical Magazine*, 19, 19–32.

Zografos, K., Ferentinos, K., & Papaioannou, T. (1989). Limiting properties of some measures of information. *Annals of the Institute of Statistical Mathematics*, 41, 451–460.

Received 18 March 2012
Accepted 08 August 2012