# Dimensionality Reduction Models in Density Estimation and Classification

**Alexander Samarov**

**Abstract** In this paper we consider the problem of multivariate density estimation assuming that the density allows some form of dimensionality reduction. Estimation of high-dimensional densities and dimensionality reduction models are important topics in nonparametric and semi-parametric econometrics.We start with the Independent Component Analysis (ICA) model, which can be considered as a form of dimensionality reduction of a multivariate density. We then consider multiple index model, describing the situations where high-dimensional data has a low-dimensional non-Gaussian component while in all other directions the data are Gaussian, and the independent factor analysis (IFA) model, which generalizes the ordinary factor analysis, principal component analysis, and ICA. For each of these models, we review recent results, obtained in our joint work with Tsybakov, Amato, and Antoniadis, on the accuracy of the corresponding density estimators, which combine model selection with estimation. One of the main applications of multivariate density estimators is in classification, where they can be used to construct plug-in classifiers by estimating the densities of each labeled class. We give a bound to the excess risk of nonparametric plug-in classifiers in terms of the MISE of the density estimators of each class. Combining this bound with the above results on the accuracy of density estimation, we show that the rate of the excess Bayes risk of the corresponding plug-in classifiers does not depend on the dimensionality of the data.

## 1 Introduction

Complex data sets lying in multidimensional spaces are a commonplace occurrence in many parts of econometrics. The need for analyzing and modeling high-dimensional data often arises in nonparametric and semi-parametric econometrics, quantitative finance, and risk management, among other areas. One of the important

A. Samarov (✉)
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
e-mail: samarov@mit.edu

J. Beran et al. (eds.), *Empirical Economic and Financial Research*,
Advanced Studies in Theoretical and Applied Econometrics 48,
DOI 10.1007/978-3-319-03122-4_30

challenges of the analysis of such data is to reduce its dimensionality in order to identify and visualize its structure.

It is well known that common nonparametric density estimators are quite unreliable even for moderately high-dimensional data. This motivates the use of dimensionality reduction models. The literature on dimensionality reduction is very extensive, and we mention here only some publications that are connected to our context and contain further references (Roweis and Saul 2000; Tenenbaum et al. 2000; Cook and Li 2002; Blanchard et al. 2006; Samarov and Tsybakov 2007).

In this paper we review several dimensionality reduction models analyzed in Samarov and Tsybakov (2004, 2007), and Amato et al. (2010).

In Sect. 2 we consider the ICA model for multivariate density where the distribution of independent sources are not parametrically specified. Following results of Samarov and Tsybakov (2004), we show that the density of this form can be estimated at one-dimensional nonparametric rate, corresponding to the independent component density with the worst smoothness.

In Sect. 3 we discuss multiple index model, describing the situations where high-dimensional data has a low-dimensional non-Gaussian component while in all other directions the data are Gaussian. In Samarov and Tsybakov (2007) we show, using recently developed methods of aggregation of density estimators, that one can estimate the density of this form, without knowing the directions of the non-Gaussian component and its dimension, with the best rate attainable when both non-Gaussian index space and its dimension are known.

In Sect. 4 we consider estimation of a multivariate density in the noisy independent factor analysis (IFA) model with unknown number of latent independent components observed in Gaussian noise. It turns out that the density generated by this model can be estimated with a very fast rate. In Amato et al. (2010) we show that, using recently developed methods of aggregation Juditsky et al. (2005, 2008), we can estimate the density of this form at a parametric root-$n$ rate, up to a logarithmic factor independent of the dimension $d$.

In Sect. 5 we give a bound to the excess risk of nonparametric plug-in classifiers in terms of the integrated mean square error (MISE) of the density estimators of each class. Combining this bound with the results of previous sections, we show that if the data in each class are generated by one of the models discussed there, the rate of the excess Bayes risk of the corresponding plug-in classifiers does not depend on the dimensionality of the data.

## 2 Nonparametric Independent Component Analysis

Independent Component Analysis (ICA) is a statistical and computational technique for identifying hidden factors that underlie sets of random variables, measurements, or signals, blind source separation. In the ICA model the observed data variables are assumed to be (linear or nonlinear) mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed non-Gaussian and mutually independent; they are called the independent components of the data.

Most of the existing ICA algorithms concentrate on recovering the mixing matrix and either assume the known distribution of sources or allow for their limited, parametric flexibility (Hyvarinen et al. 2001). Most ICA papers either use mixture of Gaussian distributions as source models or assume that the number of independent sources is known, or both. In our work, the ICA serves as a dimensionality reduction model for multivariate nonparametric density estimation; we suppose that the distribution of the sources (factors) and their number are unknown.

The standard (linear, noise-free, full rank) ICA model assumes that $d$-dimensional observations $\mathbf{X}$ can be represented as

$$\mathbf{X} = A\mathbf{U},$$

where $A$ is an unknown nonsingular $d \times d$-matrix, and $\mathbf{U}$ is an unobserved random $d$-vector with independent components. The goal of ICA is to estimate the matrix $A$, or its inverse $B^\top = A^{-1}$, based on a sample $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d. $p(\mathbf{x})$. When all components of $\mathbf{U}$, with a possible exception of one, are non-Gaussian, the mixing matrix $A$ is identifiable up to the scale and permutation of its columns.

The ICA model can be equivalently written in terms of the probability density of the observed data:

$$p(\mathbf{x}) = |\det(B)| \prod_{j=1}^{d} p_j(\mathbf{x}^\top \beta_j), \quad \mathbf{x} \in \mathbf{R}^d, \tag{1}$$

where $\beta_1, \dots, \beta_d -$ unknown, linearly independent, unit-length $d$-vectors, $\det(B)$ is the determinant of the matrix $B = (\beta_1, \dots, \beta_d)$, $B^\top = A^{-1}$, and $p_j(\cdot)$, $j = 1, \dots, d$, are probability densities of the independent sources.

Most known ICA methods specify the parametric form of the latent component densities $p_j$ and estimate $B$ together with parameters of $p_j$ using maximum likelihood or minimization of the empirical versions of various divergence criteria between densities, see, e.g., Hyvarinen et al. (2001) and the references therein. In general, densities $p_j$ are unknown, and one can consider ICA as a semiparametric model in which these densities are left unspecified.

In Samarov and Tsybakov (2004) we show that, even without knowing $\beta_1, \dots, \beta_d$, $p(\mathbf{x})$ can be estimated at one-dimensional nonparametric rate, corresponding to the independent component density with the worst smoothness. Our method of estimating $\beta_1, \dots, \beta_d$ is based on nonparametric estimation of the average outer product of the density gradient

$$T(p) = \mathbf{E}[\nabla p(X) \nabla^\top p(X)],$$

where $\nabla p$ is the gradient of $p$, and simultaneous diagonalization of this estimated matrix and the sample covariance matrix of the data. After the directions have been estimated at root-$n$ rate, the density (1) can be estimated, e.g. using the kernel estimators for marginal densities, at the usual one-dimensional nonparametric rate.

The method of Samarov and Tsybakov (2004) can be applied to a generalization of ICA where the independent components are multivariate. Our method estimates these statistically independent linear subspaces and reduces the original problem to the fundamental problem of identifying independent subsets of variables.

## 3 Multi-Index Departure from Normality Model

We consider next another important dimensionality reduction model for density:

$$p(x) \;=\; \phi_d(x) g(B^\top x), \qquad x \in \mathbf{R}^d, \tag{2}$$

where $B$—unknown $d \times m$ matrix with orthonormal columns, $1 \le m \le d$, $g : \mathbf{R}^m \to [0,\infty)$ unknown function, and $\phi_d(\cdot)$ is the density of the standard $d$-variate normal distribution.

A density of this form models the situation where high-dimensional data has a low-dimensional non-Gaussian component ($m << d$) while all other components are Gaussian. Model (2) can be viewed as an extension of the projection pursuit density estimation (PPDE) model, e.g. Huber (1985), and of the ICA model. A model similar to (2) was considered in Blanchard et al. (2006).

Note that the representation (2) is not unique. In particular, if $Q_m$ is an $m \times m$ orthogonal matrix, the density $p$ in (2) can be rewritten as $p(x) = \phi_d(x) g_1(B_1^\top x)$ with $g_1(y) = g(Q_m y)$ and $B_1 = BQ_m$. However, the linear subspace $\mathscr{M}$ spanned by the columns of $B$ is uniquely defined by (2).

By analogy with regression models, e.g. Li (1991), Hristache et al. (2001), we will call $\mathscr{M}$ the *index space*. In particular, if the dimension of $\mathscr{M}$ is 1, model (2) can be viewed as a density analog of the single index model in regression. In general, if the dimension of $\mathscr{M}$ is arbitrary, we call (2) the *multiple index model.*

When the dimension $m$ and an index matrix $B$ (i.e., any of the matrices, equivalent up to an orthogonal transformation, that define the index space $\mathscr{M}$) are specified, the density (2) can be estimated using a kernel estimator

$$\hat{p}_{m,B}(x) = \frac{\phi_d(x)}{\phi_m(B^\top x)} \frac{1}{nh^m} \sum_{i=1}^{n} K\left(\frac{B^\top (X_i - x)}{h}\right),$$

with appropriately chosen bandwidth $h > 0$ and kernel $K : \mathbf{R}^m \to \mathbf{R}^1$. One can show, see Samarov and Tsybakov (2007), that, if the function $g$ is twice differentiable, the mean integrated mean squared error (MISE) of the estimator $\hat{p}_{m,B}$ satisfies:

$$MISE(\hat{p}_{m,B}, p) := \mathbf{E}||\hat{p}_{m,B} - p||^2 = O(n^{-4/(m+4)}), \tag{3}$$

if the bandwidth $h$ is chosen of the order $h \overset{\mathbb{P}}{\sim} n^{-1/(m+4)}$. Using the standard techniques of the minimax lower bounds, it is easy to show that the rate $n^{-4/(m+4)}$ is the optimal MISE rate for this model and thus the estimator $\hat{p}_{m,B}$ with $h \overset{\mathbb{P}}{\sim} n^{-1/(m+4)}$ has the optimal rate for this class of densities.

In Samarov and Tsybakov (2007) we show, using recently developed methods of aggregation of density estimators, that one can estimate this density, without knowing $B$ and $m$, with the same rate $O(n^{-4/(m+4)})$ as the optimal rate attainable when $B$ and $m$ are known. The aggregate estimator of Samarov and Tsybakov (2007) automatically accomplishes dimension reduction because, if the unknown true dimension $m$ is small, the rate $O(n^{-4/(m+4)})$ is much faster than the best attainable rate $O(n^{-4/(d+4)})$ for a model of full dimension. This estimator can be interpreted as an adaptive estimator, but in contrast to adaptation to unknown smoothness usually considered in nonparametrics, here we deal with adaptation to unknown dimension $m$ and to the index space $\mathscr{M}$ determined by a matrix $B$.

# 4 IFA Model

In this section we consider an IFA model with unknown number and distribution of latent factors:

$$\mathbf{X} = A\mathbf{S} + \varepsilon, \tag{4}$$

where $A$ is $d \times m$ unknown deterministic matrix, $m < d$, with orthonormal columns; $\mathbf{S}$ is an $m$-dimensional random vector of independent components with unknown distributions, and $\varepsilon$ is a normal $\mathbf{N}_d(0, \sigma^2 \mathbf{I}_d)$ random vector of noise independent of $\mathbf{S}$.

By independence between the noise and the vector of factors $\mathbf{S}$, the target density $p_{\mathbf{X}}$ can be written as a convolution:

$$p_{\mathbf{X}}(\mathbf{x}) = \int_{\mathbf{R}^m} \phi_{d,\sigma^2}(\mathbf{x} - A\mathbf{s}) F_{\mathbf{S}}(d\mathbf{s}), \tag{5}$$

where $\phi_{d,\sigma^2}$ denotes the density of a $d$-dimensional Gaussian distribution $N_d(0, \sigma^2 \mathbf{I}_d)$ and $F_{\mathbf{S}}$ is the distribution of $\mathbf{S}$.

Note that (5) can be viewed as a variation of the Gaussian mixture model which is widely used in classification, image analysis, mathematical finance, and other areas, cf., e.g., Titterington et al. (1985) and McLachlan and Peel (2000). In Gaussian mixture models, the matrix $A$ is the identity matrix, $F_{\mathbf{S}}$ is typically a discrete distribution with finite support, and variances of the Gaussian terms are usually different.

Since in (5) we have a convolution with a Gaussian distribution, the density $p_{\mathbf{X}}$ has very strong smoothness properties, no matter how irregular the distribution $F_{\mathbf{S}}$

of the factors is, whether or not the factors are independent, and whether or not the mixing matrix $A$ is known. In Amato et al. (2010), we construct a kernel estimator $\hat{p}_n^*$ of $p_{\mathbf{X}}$ such that

$$\mathbf{E}||\hat{p}_n^* - p_{\mathbf{X}}||_2^2 \leq C\frac{(\log n)^{d/2}}{n}, \tag{6}$$

where $C$ is a constant and $||\cdot||_2$ is the $L_2(\mathbf{R}^d)$ norm. As in Artiles (2001) and Belitser and Levit (2001), it is not hard to show that the rate given in (6) is optimal for the class of densities $p_{\mathbf{X}}$ defined by (5) with arbitrary probability distribution $F_{\mathbf{S}}$.

Though this rate appears to be very fast asymptotically, it does not guarantee good accuracy for most practical values of $n$, even if $d$ is moderately large. For example, if $d = 10$, we have $(\log n)^{d/2} > n$ for all $n \leq 10^5$.

In order to construct our estimator, we first consider the estimation of $p_{\mathbf{X}}$ when the dimension $m$, the mixing matrix $A$, and the level of noise $\sigma^2$ are specified. Because of the orthonormality of columns of $A$, $A^\top$ is the demixing matrix: $A^\top X = S + A^\top\varepsilon$, and the density of $X$ can be written as

$$p_{\mathbf{X}}(\mathbf{x}) = \left(\frac{1}{2\pi\sigma^2}\right)^{(d-m)/2} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{x}^\top(\mathbf{I}_d - AA^\top)\mathbf{x}\right\} \prod_{k=1}^{m} g_k(\mathbf{a}_k^\top\mathbf{x}),$$

where $\mathbf{a}_k$ denotes the $k$th column of $A$ and $g_k(u) = (p_{S_k} * \phi_1)(u) = \int_R p_{S_k}(s)\phi_1(u-s)ds$.

In Amato et al. (2010) we show that, using kernel estimators for $g_k$, one can construct an estimator for the density $p_{\mathbf{X}}$ which has the mean integrated square error (MISE) of the order $(\log n)^{1/2}/n$. Note that neither $m$ nor $d$ affect the rate.

When the index matrix $A$, its rank $m$, and the variance of the noise $\sigma^2$ are all unknown, we use a model selection type aggregation procedure called the mirror averaging algorithm of Juditsky et al. (2008) to obtain fully adaptive density estimator. We make a few additional assumptions.

**Assumption 1** At most one component of the vector of factors $\mathbf{S}$ in (4) has a Gaussian distribution.

**Assumption 2** The columns of the matrix $A$ are orthonormal.

**Assumption 3** The number of factors $m$ does not exceed an upper bound $M$, $M < d$.

**Assumption 4** The $M$ largest eigenvalues of the covariance matrix $\Sigma_{\mathbf{X}}$ of the observations $\mathbf{X}$ are distinct and the 4th moments of the components of $\mathbf{X}$ are finite.

Assumption 1, needed for the identifiability of $A$, is standard in the ICA literature, see, e.g., Hyvarinen et al. (2001) Assumption 2 is rather restrictive but, as we show below, together with the assumed independence of the factors, it allows us to eliminate dependence of the rate in (6) on the dimension $d$. Assumption 3 means that model (4) indeed provides the dimensionality reduction. The assumption

$M < d$ is only needed to estimate the variance $\sigma^2$ of the noise; if $\sigma^2$ is known, we can allow $M = d$. Assumption 4 is needed to establish root-$n$ consistency of the eigenvectors of the sample covariance matrix of $\mathbf{X}$.

Under these assumptions, in Amato et al. (2010) we construct an estimator for the density of the form (5) that adapts to the unknown $m$ and $A$, i.e., has the same MISE rate $O((\log n)^{1/2}/n)$, independent of $m$ and $d$, as in the case when the dimension $m$, the matrix $A$, and the variance of the noise $\sigma^2$ are known.

# 5 Application to Nonparametric Classification

One of the main applications of multivariate density estimators is in classification, which is one of the important econometric techniques. These estimators can be used to construct nonparametric classifiers based on estimated densities from labeled data for each class.

The difficulty with such density-based plug-in classifiers is that, even for moderately large dimensions $d$, standard density estimators have poor accuracy in the tails, i.e., in the region which is important for classification purposes. In this section we consider the nonparametric classification problem and bound the excess misclassification error of a plug-in classifier in terms of the MISE of class-conditional density estimators. This bound implies that, for the class-conditional densities obeying the dimensionality reduction models discussed above, the resulting plug-in classifier has nearly optimal excess error.

Assume that we have $J$ independent training samples $\{X_{j1}, \ldots, X_{jN_j}\}$ of sizes $N_j$, $j = 1, \ldots, J$, from $J$ populations with densities $f_1, \ldots, f_J$ on $\mathbf{R}^d$. We will denote by $\mathscr{D}$ the union of training samples. Assume that we also have an observation $\mathbf{X} \in \mathbf{R}^d$ independent of these samples and distributed according to one of the $f_j$. The classification problem consists in predicting the corresponding value of the class label $j \in \{1, \ldots, J\}$. We define a classifier or prediction rule as a measurable function $T(\cdot)$ which assigns a class membership based on the explanatory variable, i.e., $T : \mathbf{R}^d \to \{1, \ldots, J\}$. The misclassification error associated with a classifier $T$ is usually defined as

$$R(T) = \sum_{j=1}^{J} \pi_j \mathbf{P}_j(T(\mathbf{X}) \neq j) = \sum_{j=1}^{J} \pi_j \int_{\mathbf{R}^d} I(T(\mathbf{x}) \neq j) f_j(\mathbf{x}) d\mathbf{x},$$

where $\mathbf{P}_j$ denotes the class-conditional population probability distribution with density $f_j$, and $\pi_j$ is the prior probability of class $j$. We will consider a slightly more general definition:

$$R_C(T) = \sum_{j=1}^{J} \pi_j \int_C I(T(\mathbf{x}) \neq j) f_j(\mathbf{x}) d\mathbf{x},$$

where $C$ is a Borel subset of $\mathbf{R}^d$. The Bayes classifier $T^*$ is the one with the smallest misclassification error:

$$R_C(T^*) = \min_T R_C(T).$$

In general, the Bayes classifier is not unique. It is easy to see that there exists a Bayes classifier $T^*$ which does not depend on $C$ and which is defined by

$$\pi_{T^*(\mathbf{x})} f_{T^*(\mathbf{x})}(\mathbf{x}) = \min_{1 \le j \le J} \pi_j f_j(\mathbf{x}), \quad \forall\, \mathbf{x} \in \mathbf{R}^d.$$

A classifier trained on the sample $\mathscr{D}$ will be denoted by $T_{\mathscr{D}}(\mathbf{x})$. A key characteristic of such a classifier is the misclassification error $R_C(T_{\mathscr{D}})$. One of the main goals in statistical learning is to construct a classifier with the smallest possible excess risk

$$\mathscr{E}(T_{\mathscr{D}}) = \mathbf{E} R_C(T_{\mathscr{D}}) - R_C(T^*).$$

We consider plug-in classifiers $\hat{T}(\mathbf{x})$ defined by:

$$\pi_{\hat{T}(\mathbf{x})} \hat{f}_{\hat{T}(\mathbf{x})}(\mathbf{x}) = \min_{1 \le j \le J} \pi_j \hat{f}_j(\mathbf{x}), \quad \forall\, \mathbf{x} \in \mathbf{R}^d$$

where $\hat{f}_j$ is an estimator of density $f_j$ based on the training sample $\{X_{j1}, \dots, X_{jN_j}\}$.

The following proposition relates the excess risk $\mathscr{E}(\hat{T})$ of plug-in classifiers to the rate of convergence of the estimators $\hat{f}_j$, see Amato et al. (2010).

**Proposition 1**

$$\mathscr{E}(\hat{T}) \le \sum_{j=1}^{J} \pi_j\, \mathbf{E} \int_C |\hat{f}_j(\mathbf{x}) - f_j(\mathbf{x})| d\mathbf{x}$$

Assume now that the class-conditional densities follow, for example, the noisy IFA model (5) with different unknown mixing matrices and that $N_j \overset{\mathbb{P}}{\sim} n$ for all $j$. Let $C$ be a Euclidean ball in $\mathbf{R}^d$ and define each of the estimators $\hat{f}_j$ using the mirror averaging procedure as in the previous section. Then, using results of that section, we have

$$\mathbf{E} \int_C |\hat{f}_j(\mathbf{x}) - f_j(\mathbf{x})| d\mathbf{x} \le \sqrt{|C|}\, \mathbf{E} \|\hat{f}_j - f_j\|_{2,C} = \mathscr{O}\left( \frac{(\log n)^{1/4}}{\sqrt{n}} \right)$$

as $n \to \infty$, where $|C|$ denotes the volume of the ball $C$ and the norm $\|\cdot\|_{2,C}$ is defined as $\|f\|_{2,C}^2 = \int_C f^2(\mathbf{x}) d\mathbf{x}$. Thus, the excess risk $\mathscr{E}(\hat{T})$ converges to 0 at the rate $(\log n)^{1/4}/\sqrt{n}$ independently of the dimension $d$.

Similarly, we can show, using the above proposition, that, if the class densities follow other dimensionality reduction models considered in this paper, the rate of the excess Bayes risk of the corresponding plug-in classifiers does not depend on the dimensionality of the data.

**Acknowledgements** Partial support provided by the Singapore-MIT Alliance in Computation and Systems Biology.

## References

Amato, U., Antoniadis, A., Samarov, A., & Tsybakov, A. (2010). Noisy independent factor analysis model for density estimation and classification. *Electronic Journal of Statistics*, *4*, 707–736.

Artiles, L. M. (2001). Adaptive Minimax Estimation in Classes of Smooth Functions (Ph.D. thesis). University of Utrecht.

Belitser, E., & Levit, B. (2001). Asymptotically local minimax estimation of infinitely smooth density with censored data. *Annals of the Institute of Statistical Mathematics*, *53*, 289–306.

Blanchard, B., Kawanabe, G. M., Sugiyama, M., Spokoiny, V., & Müller, K. R. (2006). In search of non-gaussian components of a high-dimensional distribution. *Journal of Machine Learning Research*, *7*, 247–282.

Cook, R. D., & Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics*, *32*, 455–474.

Hristache, M., Juditsky, A., Polzehl J., & Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *Annals of Statistics*, *29*, 1537–1566.

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

Hyvarinen, A., Karhunen, J., & Oja, E. (2001). *Independent component analysis*. New York: Wiley.

Juditsky, A., Rigollet, P., & Tsybakov, A. B. (2008). Learning by mirror averaging. *Annals of Statistics*, *36*, 2183–2206.

Juditsky, A. B., Nazin, A. V., Tsybakov, A. B., & Vayatis, N. (2005). Recursive aggregation of estimators by the mirror descent algorithm with averaging. *Problems of Information Transmission*, *41*, 368–384.

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

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.

Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*, 2323–2326.

Samarov, A., & Tsybakov, A. B. (2004). Nonparametric independent component analysis. *Bernoulli*, *10*, 565–582.

Samarov, A., & Tsybakov, A. B. (2007). Aggregation of density estimators and dimension reduction. In V. Nair (Ed.), *Advances in statistical modeling and inference, essays in honor of K. Doksum*. Series in Biostatistics (Vol. 3, pp. 233–251). London: World Scientific.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*, 2319–2323.

Titterington, D., Smith, A., & Makov, U. (1985). *Statistical analysis of finite mixture distributions*. New York: Wiley.