Combining eigenvalues and variation of eigenvectors for order determination

Wei Luo and Bing Li

City University of New York and Penn State University

wei.luo@baruch.cuny.edu bing@stat.psu.edu

1 Introduction of the problem

The problem Determine the rank of a matrix, or the dimension of a subspace, is very common in statistical applications. These boil down to determining how many eigenvalues of a random matrix (which estimates the target matrix) are significantly different from 0. Here are some examples.

- **Principal Component Analysis (PCA)**. Suppose we have a sample of p-dimensional random vectors $X_1, \ldots, X_n$ which are i.i.d. X. Assume

$$\text{var}(X) = \Sigma + \sigma^2 I_p$$

where Σ is a rank d matrix. At the population level, the number of principal components is the rank of Σ. At the sample level, Σ is estimated by

$$\hat{\Sigma} = E_n[(X - \bar{X})(X - \bar{X})^\intercal].$$

We need to determine how many principal components are significant. The statistical question is how to estimate the rank d of Σ based on $\hat{\Sigma}$?

- **Canonical Correlation Analysis (CCA)**. Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are an i.i.d. sample of X, Y. Suppose the covariance matrix between X and Y, say Σ_{XY}, is a singular matrix of rank $d < \min(\dim(X), \dim(Y))$. Let $\hat{\Sigma}_{XY}$ be the sample covariance matrix. In CCA we need to estimate the rank of Σ_{XY}.

- **Sufficient dimension reduction (SDR)**. Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are an i.i.d. sample of X, Y, and we would like to find $\beta \in \mathbb{R}^d$ such that

$$Y \perp\!\!\!\perp X | \beta^{\mathsf{T}} X.$$

  See Li (1991, 1992), Cook & Weisberg (1991), Cook (1994). It can be shown that

$$\Sigma_{XX}^{-1/2} E(X|Y) \in \text{span}(\beta).$$

  The sample estimator based on this is called Sliced Inverse Regression (Li, 1991).

Let J_h, $h = 1, \ldots, s$ be a partition of $\text{supp}(X)$. Let

$$E_n(X|Y) = \frac{E_n(XI(Y \in J_h))}{E_n I(Y \in J_h)}.$$

Let

$$\text{var}_s(E_n(X|Y)) = E_s[(E_n(X|Y) - E_n(X))(E_n(X|Y) - E_n(X))^\intercal]$$

Since $\text{var}_s(E_n(X|Y))$ estimates $\text{var}(E(X|Y))$, its rank is no more than d. An important problem in SDR is to decide what is the dimension off $\text{span}(\beta)$. Thus, once again, we face the task of estimating the rank of a matrix.

- Many other sufficient dimension reduction methods, such as the Sliced Average Variance Estimator (Cook & Weisberg, 1991), Contour Regression (Li, Zha & Chiaromonte, 2005), Directional Regression (Li & Wang, 2007), and Outer Product Gradient (Xia, Tong, Li & Zhu, 2002), have the similar needs for estimating the rank of a matrix from its random estimator.

The general problem extracted from these applications can be stated as follows. Suppose M is a matrix of rank d, and $\hat{M}$ is a matrix-valued statistic based on a sample $U_1, \ldots, U_n$. Moreover, suppose

$$\sqrt{n}\,\mathrm{vec}(\hat{M} - M) \xrightarrow{\mathcal{D}} N(0, \Lambda)$$

How do we estimate the rank of d?

Here and hereafter, we use U to represent either a random vector or a pair of random vectors, depending on specific applications. For example, in PCA, U is X; in CCA and SDR, $U = (X, Y)$.

Existing methods for order determination rely either on the magnitude of the eigenvalues of $\hat{M}$ or on the variability of the eigenvectors of $\hat{M}$, as outlined below.

2 Previous solutions I: scree plot

The Scree plot is the plot of eigenvalues of $\hat{M}$. The rank of M is estimated by the point where a significant drop occurs. This point is called the "elbow" of the scree

plot. The trick is how to determine if a significant drop has occurred. A typical scree plot looks like this:

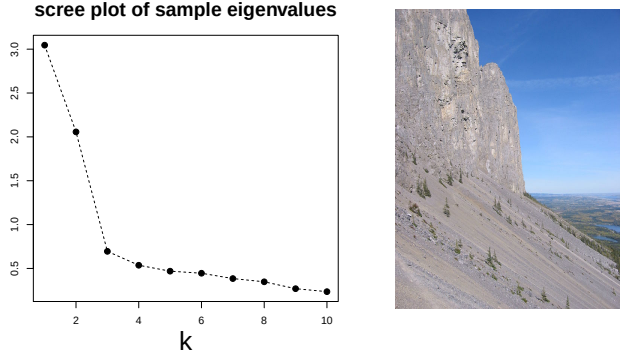


Figure 1. Scree plot and true scree

This example is based on a data set simulated from the following model:

$$Y = 0.4(X_1 + X_2 + X_3)^2 + 3\sin\{0.25(X_1 + X_5 + 3X_6)\} + 0.4\epsilon$$

where $X \sim N(0, I_p)$, $p = 10$, $\epsilon \sim N(0, 0.5^2)$, $X \perp\!\!\!\perp \epsilon$, $n = 500$. The method used to obtain the matrix $\hat{M}$ is Directional Regression (Li & Wang, 2007). In this example elbow occurs at $k = 3$; the rank of M is estimated by "eyeballing" as $\hat{d} = 2$, which is correct.

There are two types of methods that are essentially based on the scree plot.

- **Sequential test** In the sufficient dimension reduction contexts, sequential tests are developed by Li (1991), Schott (1994), Cook & Li (2004), and Bura & Yang (2011), among others. This is the basic idea. Set up a sequence of hypotheses:

  $$H_0 : \lambda_{k+1} = \cdots = \lambda_p = 0.$$

  When k is small, this hypothesis is rejected, from a certain point onwards, this hypothesis is accepted. The last k for which this hypothesis is rejected is the

estimate of the rank. Usually, the test statistic is an increasing function of the sum of small eigenvalues of $\hat{M}$, its asymptotic distribution being a linear combination of independent $\chi^2_{(1)}$.

- **Pros and Cons sequential test** Pros: clear statistical interpretation; no tuning is involved. Cons: need to make asymptotic expansions case by case; i.e. whenever you invent a new estimator you have to do the expansion. Some expansions can be quite complicated. See Li & Wang (2007). Moreover, for fixed significance level, the method is by definition not consistent.

- **BIC-type criteria** In the SDR context, this was developed in Zhu, Miao & Peng (2006), Wang & Yin (2008), Li, Artemiou & Li (2011). These criteria are based on a monotone function of the eigenvalues plus a deterministic penalty term. A version of the criteria is the following:

$$G_n(k) = \sum_{i=1}^{k} \lambda_i(\hat{M}) - c_1(n)c_2(k)$$

where $c_2(k) \uparrow k$ and $1 \succ c_1(n) \succ n^{-1/2}$. It can be shown that if $\sqrt{n}(\hat{M} - M) = O_P(1)$, then the maximizer converges to the rank of M.

- **Pros and Cons of BIC-type criteria**. Pros: easy to use; no need to compute asymptotic distributions; consistent. Cons: result heavily depends how you choose $c_1(n)$. In fact, for any particular data set, you can choose $c_1(n)$ so that the plot is maximized at *any* k! Hard to find universal $c_1(n)$ that works well for all models.

3 Previous solution II: Ye-Weiss plot

A fundamentally different idea for order determination was proposed by Ye & Weiss (2003), which are based on the variation of the orientation of eigenvectors, which goes as follows:

- **Bootstrap eigenvector variation (BEV)** Generate a set of bootstrap (Efron, 1979) samples $U_1^*, \ldots, U_n^*$, and each time estimate M by M^* based on the boot-

strap sample. For each Bootstrap sample i, and at each k, take the first k eigenvectors of M^* to form the matrix $B^*_{k,i}$. Do the same thing on the full sample to get $\hat{B}_k$. Measure the variation of $B^*_{k,i}$ around the full sample estimate $\hat{B}_k$ by

$$n^{-1}\sum_{i=1}^{n}\{1 - |\det(\hat{B}_k^{\mathsf{T}} B^*_{k,i})|\}.$$

Each term in the summand is near 0 if $\hat{B}_k$ and $B^*_{k,i}$ are close; near 1 if they are nearly orthogonal. Intuitively, before k reaches d, all the eigenvalues are distinct, so the eigenvectors of M^* and $\hat{M}$ converge to fixed vectors, and the criterion should be small. After k reaches d, all the eigenvalues are 0, so the eigenvalues of M in the null space are arbitrary. So there is nothing fixed for $\hat{M}$ and M^* to converge to, and the above criterion should be large. Indeed, a typical plot of the above criterion against k looks like this (generated from the same model):

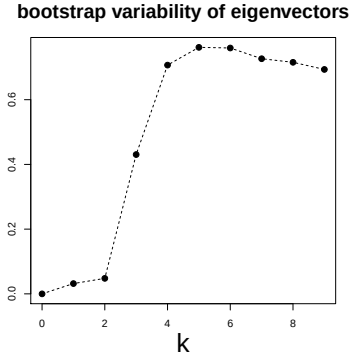


Figure 2. Typical bootstrap eigenvector variation plots

We can see the curve has a sharp upturn after reaching 2.

- **Pros and Cons of BEV.** Pros: no need to derive asymptotic distribution. No arbitrary tuning constant. Cons. Patterns not so clear cut when some of

the nonzero eigenvalues of M are also the same. First, consider the eigenvalue pattern

4,3,2,1,0,0,0,...

In this case, the BEV pattern would be

small, small, small, small, large, large,...

as shown in the above plot.

But, now consider the eigenvalue pattern

3,3,3,2,2,2,0,0,0, ...

Then the bootstrap BEV pattern would be:

large, large, small, large, large, small, large, large, large, ...

4 New method: the ladle plot

The above discussion reveals that both eigenvalues and the eigenvectors of $\hat{M}$ contain useful information about the rank d of M. Specifically:

> **When the eigenvalues of a random matrix are far apart, the bootstrap variability of the corresponding eigenvectors tends to be small; when the eigenvalues are close together, this bootstrap variability tends to be large. Thus, by juxtaposing these two quantities, we can pinpoint the rank of M more accurately than using either one of them alone!**

Based on this observation, we introduce the following functions

- Rescaled scree plot:

$$\phi_n : \{0, \ldots, p-1\} \to \mathbb{R}, \quad \phi_n(\lambda) = \hat{\lambda}_{k+1} / (\textstyle\sum_{i=1}^{p} \hat{\lambda}_i)$$

- Bootstrap eigenvector variability:

$$f_n^0(k) = \begin{cases} 0, & k = 0, \\ n^{-1} \sum_{i=1}^{n} \{1 - |\det(\hat{B}_k^{\intercal} B_{k,i}^*)|\}, & k = 1, \ldots, p-1. \end{cases}$$

- We rescaled bootstrap eigenvector variability:

$$f_n : \{0, \ldots, p-1\} \to \mathbb{R}, \quad f_n(k) = f_n^0(k) \,/\, \textstyle\sum_{i=0}^{p-1} f_n^0(i).$$

- Scale-free combination:

$$g_n : \{0, \ldots, p-1\} \to \mathbb{R}, \quad g_n(k) = f_n(k) + \phi_n(k);$$

this collects information from both eigenvectors and eigenvalues. This combination is scale free - makes sense because the two quantities combine are of different natures.

- Using the same model, the same method, the same data, the next plot shows the shape of $g_n(k)$. It looks like a ladle with handle pointing to the right, and bottom at the estimator of the rank of M.

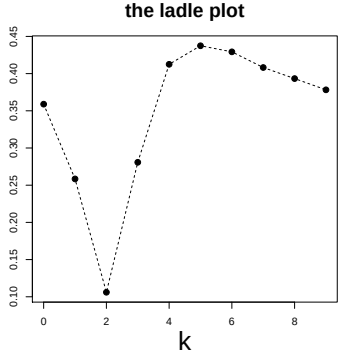
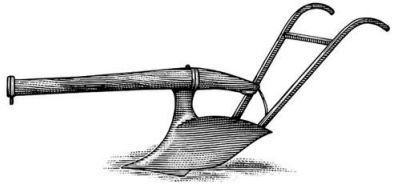


Figure 3. The ladle plot and true ladle

We experimented with many simulated and real data, and this ladle shape is quite typical. So we call it the ladle plot.

My student (first author Wei Luo) wants to call it the "plow plot"!



5 Asymptotic theory of the ladle plot

Can we show mathematically the bottom of ladle converges to the rank of M? Yes. Define some notations:

- $S = \{(X_1, Y_1), (X_2, Y_2), \ldots\}$ be a sequence of independent copies of (X, Y);
- F be the distribution of (X, Y);
- F_n be the empirical distribution based on $(X_1, X_1), \ldots, (X_n, Y_n)$;
- Given S, let $(X^*_{1,n}, Y^*_{1,n}), \ldots, (X^*_{n,n}, Y^*_{n,n})$ be i.i.d. sample from F_n;
- Given S, F^*_n the empirical distribution of $(X^*_{1,n}, Y^*_{1,n}), \ldots, (X^*_{n,n}, Y^*_{n,n})$;

In the bootstrap context, there are two layers of randomness. First, the sequence S is drawn from the true distribution of (X, Y). Second, given the sequence S, the bootstrap sample is drawn from F_n. Let $\hat{d}$ be the bottom of ladle, then it, too, has

two layers of randomness. Then in what sense can we talk about convergence of $\hat{d}$? Answer:

> **convergence (in probability, in distribution, and so on) given S for almost every S**

See Bickel & Freedman (1981).

We make the following assumption:

- Asymptotic linearity. There is a random matrix $H(X, Y)$ with mean 0 and finite second moment such that

$$\hat{M} = M + E_n H(X, Y) + o_P(n^{-1/2}). \tag{1}$$

  This assumption implies

$$n^{1/2}\{\text{vech}(\hat{M}) - \text{vech}(M)\} \to N\left(0, \text{var}_F[\text{vech}\{H(X, Y)\}]\right)$$

where $\to$ means convergence in distribution.

This condition is quite mild. For example, a sufficient condition is the statistic is Frechet differentiable with respect to the underlying distribution in a space of distributions with supremum norm. This requirement is satisfied by PCA, CCA, and most of the SDR methods, which are nice functions of sample moments.

- Self similarity The bootstrap estimator M^* satisfies

$$n^{1/2}\{\text{vech}(M^*) - \text{vech}(\hat{M})\} \rightsquigarrow N\left(0, \text{var}_F[\text{vech}\{H(X,Y)\}]\right)$$

where $\rightsquigarrow$ means convergence in distribution given S, for almost every S. Intuitively, this assumption means that the relation between M^* and $\hat{M}$ is similar to the relation between $\hat{M}$ and M; so resampling is like taking a selfie! This condition is proved for many statistics; see Bickel & Freedman (1981), Parr (1985), Liu, Singh & Lo (1989) and Gill (1989). Here we take if for granted.

- Uniform integrability. For any sequence of nonnegative random variables $\{Z_n : n = 1, 2, \ldots\}$ involved hereafter, if $Z_n = O_P(c_n)$ for some sequence $\{c_n : n \in \mathbb{N}\}$ with $c_n > 0$, then $E(c_n^{-1} Z_n)$ exists for each n and $E(c_n^{-1} Z_n) = O(1)$.

The next theorem is the mathematical characterization of BEV.

Theorem 1 *Let $c_n = [\log\{\log(n)\}]^{-2}$. Under the above three assumptions, for any positive semi-definite matrix $M \in \mathbb{R}^{p \times p}$ of rank $d < p - 1$, and $k < p$, the following relation holds for almost every sequence $S = \{(X_n, Y_n) : n = 1, 2, \ldots\}$:*

$$f_n(k) = \begin{cases} O_P(c_n^{-1} n^{-1}), & \lambda_k > \lambda_{k+1}, \\ O_P(c_n), & \lambda_k = \lambda_{k+1}, \end{cases}$$

where the probability in O_P and o_P is the conditional probability given a sequence $S = \{(X_n, Y_n) : n = 1, 2, \ldots\}$.

The point of this theorem is c_n is much larger than $c_n^{-1} n^{-1}$. In fact, $[\log(\log(n))]^{-2}$ is almost a constant. Using this result we can prove the following consistency result.

Theorem 2 *Under the three assumptions, for any positive semi-definite matrix* $M \in \mathbb{R}^{p\times p}$ *of rank* $d < p-1$*, the ladle estimator* $\hat{d}$ *satisfies:*

$$\mathrm{pr}\{\lim_{n\to\infty} \mathrm{pr}(\hat{d} = d \,|\, S) = 1\} = 1. \tag{2}$$

A special case is when all the nonzero eigenvalues of M are equal, in which Theorem 1 suggests a ladle shape of f_n itself. By applying the proof of Theorem 2, one can easily show that the minimizer of f_n is also consistent in this special case. In this sense, Theorem 2 covers the consistency of the minimizer of the bootstrap criterion in Ye & Weiss (2003) as a special case.

6 Simulation studies

We conducted simulation studies to compare the ladle estimator with existing order-determination methods, including:

- Sequential test (Bura & Yang, 2011). Two versions: the scaled version of the weighted chi-square test and the Wald-type chi-square test;
- BIC-type criterion in Zhu, Miao & Peng (2006), using their recommended tuning constants and number of slides;
- Ye-Weiss procedure (Ye & Weiss, 2003). Using three thresholds for increase in BEV.

For comprehensiveness we used twelve models, many of them taken from published papers:
1. $Y = X_1 \,/\, \{0.5 + (1.5 + X_2)^2\} + \epsilon$;
2. $Y = \exp(X_1)\, I(\epsilon > 0) - \exp(X_2)\, I(\epsilon < 0)$;
3. $Y = X_1\,(X_1 + X_2 + 1) + \epsilon$;
4. $Y = \cos(X_1) + \cos(2\,X_2) + \epsilon$;
5. $Y = X_1^2 + X_2^2 + \epsilon$;
6. $Y = X_1^2 + (X_2 + X_3)^2\,\epsilon$;

7. $Y = X_1^2 + X_2 + \epsilon$;
8. $Y = 0.4\,(X_1 + X_2 + X_3)^2 + 3\,\sin\{0.25\,(X_1 + X_5 + 3\,X_6)\} + 0.4\,\epsilon$;
9. $Y = 3\,\sin\{0.25\,(X_1 + X_2 + X_3)\} + 3\,\sin\{0.25\,(X_1 + X_5 + 3\,X_6)\} + 0.4\,\epsilon$;
10. $Y = X_1 + \epsilon$;
11. $Y = \cos(X_1) + \epsilon$;
12. $Y = 3\sin\{0.25(X_1 + X_2 + X_3)\} + \epsilon$.

In all cases, $X \sim N(0, I_p)$, $\epsilon \sim N(0, 0.5^2)$, and $X \perp\!\!\!\perp \epsilon$. The reported numbers are the percentage of correct estimation in 500 simulated samples.

Models	ST1	ST2	BIC	YW1	YW2	YW3	LADLE
1	76	50	49	97	98	75	99
2	78	34	31	100	97	64	100
3	81	37	73	61	94	80	91
4	12	60	3	33	66	93	90
5	98	37	7	99	99	83	99
6	37	4	65	83	97	90	99
7	49	92	5	55	88	83	99
8	83	96	27	95	99	74	100
9	84	46	13	4	39	66	80
10	84	66	100	100	97	55	100
11	50	91	0	100	99	80	100
12	84	0	3	100	99	65	100

We see that the ladle estimator was clearly the best overall performer. For most of the 12 models, the sequential tests have not approached their nominal level of

0.05 at the current sample size. The Bayesian information criterion was highly model dependent. The bootstrap procedures performed well in most cases, among which the top candidate had threshold $\delta = 0.4$. However, as Model 4 and Model 9 suggested, this threshold was not universally optimal.

7 Wine cultivars identification

We now apply the ladle estimator in conjunction with directional regression to the wine cultivar data in Forina, Leardi, Armanino & Lanteri (1988), which consist of 178 wine samples from three different cultivars. For each sample, its cultivar and 13 covariates are recorded, which include alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, oroanthocyanins, color intensity, hue, OD280/OD315 of diluted wines and proline. The goal is to identify the cultivars based on the covariates. We use directional regression to find a low-dimensional predictor and identify the cultivars in the reduced space. The order-

determination methods would then tell us how many sufficient predictors we should take from directional regression. To satisfy the requirements by directional regression on the covariates, we take logarithm on Malic acid, Color intensity and Proline, and take reciprocal of Magnesium, and then standardize each component of the modified predictor. For details about these requirements, see, for example, Li & Wang (2007).

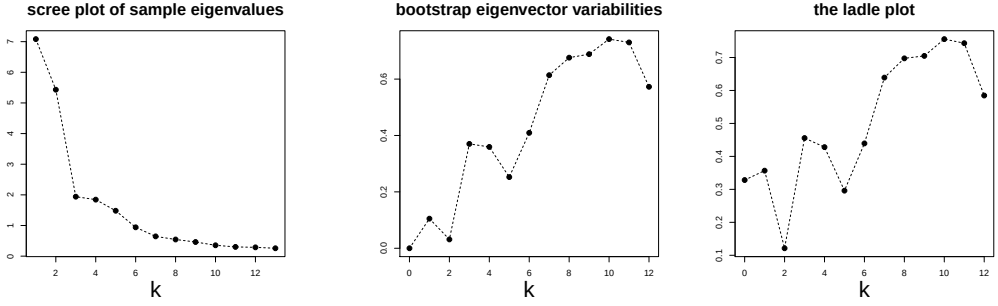


Figure 4. Scree plot, BEV plot, and ladle plot for wine cultivar data

Figure 4 shows the scree plot, BEV plot, and ladle plot. The bootstrap procedure estimates d to be 2, 5 and 6 for $\delta = 0.2, 0.4$ and 0.6, respectively; the ladle estimator yields $\hat{d} = 2$. Figure 5 is the scatter plot of the first two predictors found by directional regression, $(\hat{\beta}_1^{\mathsf{T}} X, \hat{\beta}_2 X)$, standardized to have mean 0 and identity variance matrix.

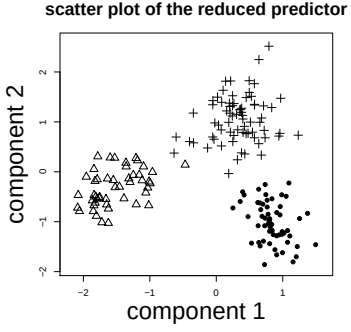


Figure 5. Scatter plot for the first two predictors of Directional Regression: three plotting symbols representing three cultivars.

To see that $\hat{d} = 2$ is a reasonable estimate, we note that the first two directional regression predictors give nearly perfect separation of the three cultivars, suggesting that two directions are sufficient to identify the cultivars; that is, the dimension of the central subspace is less than or equal to 2. In the meantime, both axes in the scatter plot offer significant separations of the three groups, suggesting that the dimension d is at least 2. These indicate that $\hat{d} = 2$ is a good choice.

References

Bickel, P. & Freedman, D. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics* **9**, 1196–1217.

Bura, E. & Yang, J. (2011). Dimension estimation in sufficient dimension reduction: a unifying approach. *Journal of Multivariate Analysis* **102**, 130–142.

Cook, R. D. (1994). Using dimension reduction subspaces to identify important imputs in models of physical systems. *In 1994 Proceedings of the Section on Physical*

and Engineering Sciences: American Statistical Association, Alexandria, VA., , 18–25.

Cook, R. D. & Li, B. (2004). Determining the dimension of iterative hessian transformation. *The Annals of Statistics* **32**, 2501–2531.

Cook, R. D. & Weisberg, S. (1991). Discussion of "sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**, 316–342.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7**, 1–26.

Forina, M., Leardi, R., Armanino, C. & Lanteri, S. (1988). *PARVUS - An extendable package of programs for data exploration, classification and correlation.* Elsevier.

Gill, R. (1989). Non- and semi-parametric maximum likelihood estimators and the von mises method, part 1. *Scandinavian Journal of Statistics* **16**, 97–128.

Li, B., Artemiou, A. & Li, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *The Annals of Statistics* **39**, 3182–3210.

Li, B. & Wang, S. (2007). On directional regression for dimension reduction. *Journal of American Statistical Association* **102**, 997–1008.

Li, B., Zha, H. & Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics* **33**, 1580–1616.

Li, K. (1992). On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma. *Journal of the American Statistical Association* **87**, 1025–1039.

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

Liu, R. Y., Singh, K. & Lo, S. H. (1989). On a representation related to the bootstrap. *Sankhyā: The Indian Journal of Statistics* **51**, 168–177.

Parr, W. (1985). The bootstrap: some large sample theory and connections with robustness. *Statistics & Probability Letters* **3**, 97–100.

Schott, J. R. (1994). Determining the demensionality in sliced inverse regression. *Journal American Statistical Association* **89**, 141–148.

Wang, Q. & Yin, X. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: sparse mave. *Computational Statistics and Data Analysis* **52**, 4512–4520.

Xia, Y., Tong, H., Li, W. K. & Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society. Series B* **64**, 363–410.

Ye, Z. & Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association* **98**, 968–979.

Zhu, L., Miao, B. & Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association* **101**, 630–642.