SAMPLE SIZE COGNIZANT DETECTION OF SIGNALS IN WHITE NOISE

Raj Rao Nadakuditi[*]

Massachusetts Institute of Technology
Department of EECS
Cambridge, MA 02139.

Alan Edelman

Massachusetts Institute of Technology
Department of Mathematics
Cambridge, MA 02139.

ABSTRACT

The detection and estimation of signals in noisy, limited data is a problem of interest to many scientific and engineering communities. We present a computationally simple, sample eigenvalue based procedure for estimating the number of high-dimensional signals in white noise when there are relatively few samples. We highlight a fundamental asymptotic limit of sample eigenvalue based detection of weak high-dimensional signals from a limited sample size and discuss its implication for the detection of two closely spaced signals.

This motivates our heuristic definition of the *effective number of identifiable signals*. Numerical simulations are used to demonstrate the consistency of the algorithm with respect to the effective number of signals and the superior performance of the algorithm with respect to Wax and Kailath's "asymptotically consistent" MDL based estimator.

Index Terms— Signal detection, eigen-inference, random matrices

1. INTRODUCTION

The observation vector, in many signal processing applications, can be modelled as a superposition of a finite number of signals embedded in additive noise. Detecting the number of signals present becomes a key issue and is often the starting point for the signal parameter estimation problem. When the signals and the noise are assumed to be samples of a stationary, ergodic Gaussian vector process, the sample covariance matrix formed from m observations has the Wishart distribution.

The proposed algorithm uses an information theoretic criterion, motivated by the approach taken by Wax and Kailath (henceforth WK) in [1], for determining the number of signals in white noise by performing inference on the eigenvalues of the resulting sample covariance matrix. The form of the estimator is motivated by the distributional properties of moments of the eigenvalues of large dimensional Wishart matrices [2].

The proposed estimator was derived by explicitly accounting for the blurring and fluctuations of the eigenvalues due to sample size constraints. Consequently, there is a greater theoretical justification for employing the proposed estimator in sample starved settings unlike the WK estimators which were derived assuming that the sample size greatly exceeds the number of sensors. This is reflected in the improved performance relative to the "asymptotically consistent" WK MDL based estimator.

Another important contribution of this paper is the description of a fundamental limit of eigen-inference, *i.e.*, inference using the sample eigenvalues alone. The concept of *effective number of identifiable signals*, introduced herein, explains why, asymptotically, if the signal level is below a threshold that depends on the noise variance, sample size and the dimensionality of the system, then reliable detection is not possible.

This paper is organized as follows. The problem is formulated in Section 2. An estimator for the number of signals present that exploits results from random matrix theory is derived in Section 3. The fundamental limits of sample eigenvalue based detection and the concept of *effective number of signals* are discussed in Section 4. Simulation results are presented in Section 5 while some concluding remarks and directions for future research are presented in Section 6.

2. PROBLEM FORMULATION

We observe m samples ("snapshots") of possibly signal bearing n-dimensional snapshot vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$ where for each i, $\mathbf{x}_i \sim \mathcal{N}_n(0, \mathbf{R})$ and $\mathbf{x}_i$ are mutually independent. The snapshot vectors are modelled as

$$\mathbf{x}_i = \mathbf{A}\,\mathbf{s}_i + \mathbf{z}_i \qquad \text{for } i = 1, \ldots, m, \qquad (1)$$

where $\mathbf{z}_i \sim \mathcal{N}_n(0, \sigma^2\mathbf{I})$, denotes an n-dimensional (real or complex) Gaussian noise vector where σ^2 is generically unknown, $\mathbf{s}_i \sim \mathcal{N}_k(\mathbf{0}, \mathbf{R}_s)$ denotes a k-dimensional (real or complex) Gaussian signal vector with covariance $\mathbf{R}_s$, and $\mathbf{A}$ is a $n \times k$ unknown non-random matrix.

Since the signal and noise vectors are independent of each other, the covariance matrix of $\mathbf{x}_i$ can hence be decomposed as

$$\mathbf{R} = \mathbf{\Psi} + \sigma^2\mathbf{I} \qquad (2)$$

where

$$\mathbf{\Psi} = \mathbf{A}\mathbf{R}_s\mathbf{A}', \qquad (3)$$

with $'$ denoting the conjugate transpose. Assuming that the matrix $\mathbf{A}$ is of full column rank, *i.e.*, the columns of $\mathbf{A}$ are linearly independent, and that the covariance matrix of the signals $\mathbf{R}_s$ is nonsingular, it follows that the rank of $\mathbf{\Psi}$ is k. Equivalently, the $n - k$ smallest eigenvalues of $\mathbf{\Psi}$ are equal to zero.

If we denote the eigenvalues of $\mathbf{R}$ by $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ then it follows that the smallest $n - k$ eigenvalues of $\mathbf{R}$ are all equal to σ^2 so that

$$\lambda_{k+1} = \lambda_{k+2} = \ldots = \lambda_n = \lambda = \sigma^2. \qquad (4)$$

Thus, if the true covariance matrix $\mathbf{R}$ were known apriori, the dimension of the signal vector k can be determined from the multiplicity of the smallest eigenvalue of $\mathbf{R}$. The problem in practice is that the covariance matrix $\mathbf{R}$ is unknown so that such a straight-forward algorithm cannot be used. The signal detection and estimation problem is hence posed in terms of an inference problem on m samples of n-dimensional multivariate real or complex Gaussian snapshot vectors.

A classical approach to this problem, developed by Bartlett [3] and Lawley [4], uses a sequence of hypothesis tests. Though this approach is sophisticated, the main problem is the subjective judgement needed by the practitioner in selecting the threshold levels for the different tests. This was overcome by Wax and Kailath in [1] wherein they propose an estimator for the number of signals (assuming $m > n$) based on the eigenvalues $l_1 \geq l_2 \geq \ldots \geq l_n$ of the sample covariance matrix (SCM) defined by

$$\widehat{\mathbf{R}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i' = \frac{1}{m}\mathbf{X}\mathbf{X}' \qquad (5)$$

where $\mathbf{X} = [\mathbf{x}_1 | \ldots | \mathbf{x}_m]$ is the matrix of observations (samples). The Akaike Information Criteria (AIC) form of the estimator is given by

$$\hat{k}_{\mathrm{AIC}} = \operatorname*{arg\,min}_{k \in \mathbb{N}:0 \leq k<n} -2(n-k)m \log \frac{g(k)}{a(k)} + 2k(2n-k) \qquad (6)$$

while the Minimum Descriptive Length (MDL) criterion is given by

$$\hat{k}_{\mathrm{MDL}} = \operatorname*{arg\,min}_{k \in \mathbb{N}:0 \leq k<n} -(n-k)m \log \frac{g(k)}{a(k)} + \frac{1}{2}k(2n-k) \log m \qquad (7)$$

where $g(k) = \prod_{j=k+1}^{n} l_j^{1/(n-k)}$ is the geometric mean of the $n - k$ smallest sample eigenvalues and $a(k) = \frac{1}{n-k}\sum_{j=k+1}^{n} l_j$ is their arithmetic mean.

It is known [1] that the AIC form inconsistently estimates the number of signals, while the MDL form estimates the number of signals consistently. The simplicity of the estimator, and the large sample consistency are among the primary reasons why the Kailath-Wax MDL estimator continues to be employed in practice [5]. In the two decades since the publication of the WK paper, researchers have come up with many innovative solutions ([6, 7, 8] to list a few) for making the estimators more robust by exploiting some type of prior knowledge.

The most important deficiency of the WK and related estimators that remains unresolved occurs when the sample size is smaller than the number of sensors, *i.e.*, when $m < n$. In this situation, the SCM is singular and the estimators become degenerate. Practitioners often overcome this in an ad-hoc fashion by, for example, restricting k in (7) to integer values in the range $0 \leq k < \min(n, m)$. Since large sample, *i.e.*, $m \gg n$, asymptotics were used to derive the estimators in [1], there is no rigorous theoretical justification for such a reformulation even if the simulation results suggest that the WK estimators are working "well enough."

Other sample eigenvalue based solutions found in the literature that exploit the sample eigenvalue order statistics [6], or employ a Bayesian framework by imposing priors on the number of signals [9] are computationally more intensive and do not address the sample size starved setting in their analysis or their simulations. Particle filter based techniques [8], while useful, require the practitioner to the model the eigenvectors of the underlying population covariance matrix as well; this makes them especially sensitive to model mismatch errors that are endemic to high-dimensional settings. This motivates our development of an sample eigenvalue based estimator with a computational complexity comparable to that of the WK estimators.

3. ESTIMATING THE NUMBER OF SIGNALS

Given an observation $\mathbf{y} = [y(1), \ldots, y(N)]$ and a family of models, or equivalently a parameterized family of probability densities $f(\mathbf{y}|\boldsymbol{\theta})$ indexed by the parameter vector $\boldsymbol{\theta}$, we select the model which gives the minimum Akaike Information Criterion (AIC) [10] defined by

$$\mathrm{AIC}_k = -2\log f(\mathbf{y}|\widehat{\boldsymbol{\theta}}) + 2k \qquad (8)$$

where $\widehat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$, and k is the number of free parameters in $\boldsymbol{\theta}$. We derive an AIC based estimator for the number of signals by exploiting the following distributional properties of the moments of eigenvalues of the (signal-free) SCM.

Theorem 1 *(Dumitriu-Edelman [2]) Assume $\widehat{\mathbf{R}}$ is formed from m snapshots modelled as (1) with $k = 0$, $\lambda = 1$ then as $m, n \to \infty$ and $c_m = n/m \to c \in (0, \infty)$, then*

$$\begin{bmatrix} \sum_{i=1}^{n} l_i - n \\ \sum_{i=1}^{n} l_i^2 - n\,(1+c) - (\frac{2}{\beta}-1)c \end{bmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\,(\mathbf{0}, \mathbf{Q})$$

where $\mathcal{D}$ denotes convergence in distribution, $\beta = 1$ (or 2) when $\mathbf{x}_i$ is real (or complex) valued, respectively, and

$$\mathbf{Q} = \frac{2}{\beta} \begin{bmatrix} c & 2c\,(c+1) \\ 2c\,(c+1) & 2c\,(2c^2 + 5c + 2) \end{bmatrix}.$$

Proposition 2 *Assume $\widehat{\mathbf{R}}$ satisfies the hypotheses of Theorem 1 for some λ then as $m, n \to \infty$ and $c_m = n/m \to c \in (0, \infty)$, then*

$$n\,[t_n - (1+c)] \xrightarrow{\mathcal{D}} \mathcal{N}\left(\left(\frac{2}{\beta}-1\right)c, \frac{4}{\beta}c^2\right) \qquad (9)$$

and the test statistic t_n is given by

$$t_n = \frac{\frac{1}{n}\sum_i l_i^2}{\left(\frac{1}{n}\sum_i l_i\right)^2} = \frac{\text{Second moment of eigs}}{\text{Mean sq. of eigs}}$$

PROOF. This follows from applying the delta method [11] to the results in Theorem 1. $\square$

When $k > 0$ signals are present and assuming $k \ll n$, then the distributional properties of the $n - k$ "noise" eigenvalues are closely approximated by the distributional properties of the eigenvalues given by Theorem 1 of the signal-free SCM, *i.e.*, $k = 0$. Hence, by evaluating the statistic in Proposition 2 over a sliding window, and using the normal approximation for the statistic from Proposition 2 with $c \approx n/m$ and $k + 1$ free parameters in the AIC formulation in (8) results in the estimator:

$$\hat{k}_{\mathrm{NEW}} = \underset{k \in \mathbb{N}: 0 \le k < \min(n,m)}{\arg\min} \left\{ \frac{\beta}{4}\left[\frac{m}{n}\right]^2 q_k^2 \right\} + 2(k+1),$$

$$(10\mathrm{a})$$

where

$$q_k = n\left[\underbrace{\frac{\frac{1}{n-k}\sum_{i=k+1}^{n} l_i^2}{(\frac{1}{n-k}\sum_{i=k+1}^{n} l_i)^2}}_{t_{n,k}} - \left(1 + \frac{n}{m}\right) \right] -$$

$$\left(\frac{2}{\beta}-1\right)\frac{n}{m}. \qquad (10\mathrm{b})$$

Here $\beta = 1$ if $\mathbf{x}_i \in \mathbb{R}^n$, and $\beta = 2$ if $\mathbf{x}_i \in \mathbb{C}^n$. When the measurement vectors represent quaternion valued narrowband signals, then we set $\beta = 4$. Quaternion valued vectors arise when the data collected from vector sensors is represented using quaternions as in [12].

4. FUNDAMENTAL LIMIT OF DETECTION

The following result exposes when the "signal" eigenvalues are asymptotically distinguishable from the "noise" eigenvalues.

Proposition 3 *Assume $\widehat{\mathbf{R}}$ satisfies the hypotheses of Theorem 1. Denote the eigenvalues of $\mathbf{R}$ by $\lambda_1 \ge \lambda_2 > \ldots \ge \lambda_k > \lambda_{k+1} = \ldots \lambda_n = \lambda = \sigma^2$. Let l_j denote the j-th largest eigenvalue of $\widehat{\mathbf{R}}$. Then as $n, m \to \infty$ with $c_m = n/m \to c \in (0, \infty)$, and $j = 1, \ldots, k+1$,*

$$l_j \to \begin{cases} \lambda_j\left(1 + \dfrac{\sigma^2 c}{\lambda_j - \sigma^2}\right) & \text{if } \lambda_j > \sigma^2\,(1+\sqrt{c}) \\[2mm] \sigma^2\,(1+\sqrt{c})^2 & \text{if } \lambda_j \le \sigma^2(1+\sqrt{c}) \end{cases}$$

$$(11)$$

where the convergence is almost surely.

PROOF. This result appears in [13] for very general settings. A matrix theoretic proof for when $c < 1$ for the real case may be found in [14] and an interacting particle system interpretation appears in [15]. $\square$

Motivated by Proposition 3, we define the effective number of signals as

$$k_{\mathrm{eff}}(\mathbf{R}) = \#\ \text{eigs. of } \mathbf{R} > \sigma^2\left(1 + \sqrt{\frac{n}{m}}\right). \qquad (12)$$

4.1. Identifiability of closely spaced signals

Suppose there are two uncorrelated (hence, independent) signals so that $\mathbf{R}_s = \text{diag}(\sigma_{S1}^2, \sigma_{S2}^2)$. In (1), let $\mathbf{A} = [\mathbf{v}_1 \mathbf{v}_2]$. In a sensor array processing application, we think of $\mathbf{v}_1 \equiv \mathbf{v}(\theta_1)$ and $\mathbf{v}_2 \equiv \mathbf{v}_2(\theta_2)$ as encoding the array manifold vectors for a source and an interferer with powers σ_{S1}^2 and σ_{S2}^2, located at θ_1 and θ_2, respectively. The covariance matrix given by

$$\mathbf{R} = \sigma_{S1}^2 \mathbf{v}_1 \mathbf{v}_1' + \sigma_{S2}^2 \mathbf{v}_2 \mathbf{v}_2' + \sigma^2 \mathbf{I} \qquad (13)$$

has the $n - 2$ smallest eigenvalues $\lambda_3 = \ldots = \lambda_n = \sigma^2$ and the two largest eigenvalues

$$\lambda_1 = \sigma^2 + \frac{(\sigma_{S1}^2 \|\mathbf{v}_1\|^2 + \sigma_{S2}^2 \|\mathbf{v}_2\|^2)}{2}$$
$$+ \frac{\sqrt{\left(\sigma_{S1}^2 \|\mathbf{v}_1\|^2 - \sigma_{S2}^2 \|\mathbf{v}_2\|^2\right)^2 + 4\sigma_{S1}^2 \sigma_{S2}^2 |\langle \mathbf{v}_1, \mathbf{v}_2 \rangle|^2}}{2} \qquad (14\text{a})$$

$$\lambda_2 = \sigma^2 + \frac{(\sigma_{S1}^2 \|\mathbf{v}_1\|^2 + \sigma_{S2}^2 \|\mathbf{v}_2\|^2)}{2}$$
$$- \frac{\sqrt{\left(\sigma_{S1}^2 \|\mathbf{v}_1\|^2 - \sigma_{S2}^2 \|\mathbf{v}_2\|^2\right)^2 + 4\sigma_{S1}^2 \sigma_{S2}^2 |\langle \mathbf{v}_1, \mathbf{v}_2 \rangle|^2}}{2} \qquad (14\text{b})$$

respectively. Applying the result in Proposition 3 allows us to express the effective number of signals as

$$k_{\text{eff}} = \begin{cases} 2 & \text{if} \quad \sigma^2 \left(1 + \sqrt{\frac{n}{m}}\right) < \lambda_2 \\[2em] 1 & \text{if} \quad \lambda_2 \leq \sigma^2 \left(1 + \sqrt{\frac{n}{m}}\right) < \lambda_1 \\[2em] 0 & \text{if} \quad \lambda_1 \leq \sigma^2 \left(1 + \sqrt{\frac{n}{m}}\right) \end{cases} \qquad (15)$$

In the special situation when $\| \mathbf{v}_1 \| = \| \mathbf{v}_2 \| = \| \mathbf{v} \|$ and $\sigma_{S1}^2 = \sigma_{S2}^2 = \sigma_S^2$, we can (in an asymptotic sense) reliably detect the presence of *both signals* from the sample eigenvalues alone whenever

$$\boxed{\sigma_S^2 \| \mathbf{v} \|^2 \left(1 - \frac{|\langle \mathbf{v}_1, \mathbf{v}_2 \rangle|}{\| \mathbf{v} \|}\right) > \sigma^2 \sqrt{\frac{n}{m}}} \qquad (16)$$

Equation (16) captures the tradeoff between the identifiability of two closely spaced signals, the dimensionality of the system, the number of available snapshots and the cosine of the angle between the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. It may prove to be a useful heuristic for experimental design.
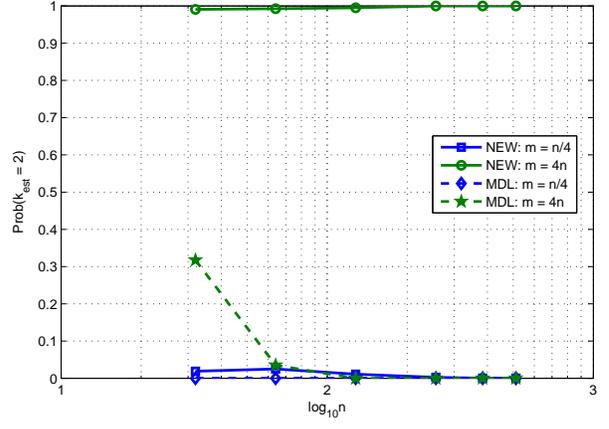
(a) $\text{Prob}(\hat{k} = 2)$ versus n for fixed n/m.
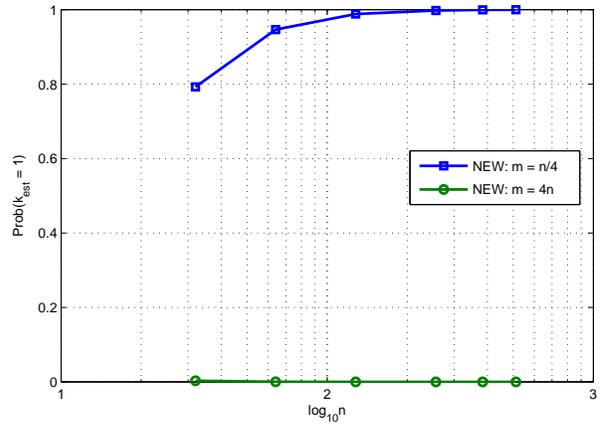
(b) $\text{Prob}(\hat{k} = 1)$ versus n for fixed n/m.

Fig. 1. Comparison of the estimators over 20,000 trials.

5. SIMULATIONS

Assume the covariance matrix $\mathbf{R}$ has $n-2$ "noise" eigenvalues with $\sigma^2 = 1$, and two "signal" eigenvalues with $\lambda_1 = 10$ and $\lambda_2 = 3$. When $m = 4n$ samples are available, Figure 1(a) shows that the proposed estimator consistently detects two signals while the WK MDL estimator does not. However, when $m = n/4$, Figure 1(a) suggests that neither estimator is able to detect both the signals present. A closer examination of the empirical data presents a different picture. For the covariance matrix considered, when $m = n/4$, then from (12), $k_{eff} = 1$. Figure 1(b) shows that for large n and $m = n/4$, the new estimator consistently estimates one signal, as expected. The WK MDL estimator detects no signals. We conjecture that the new estimator consistently estimates k_{eff} in the $n, m \to \infty, n/m \to c$ sense.

6. CONCLUSIONS

An estimator for the number of signals in white noise was presented that exhibits robustness to high-dimensionality, and sample size constraints. The concept of *effective number of signals* described provides insight into the (asymptotic) regime in which reliable detection with sample eigenvalue based methods, including the proposed method, is possible. This helps identify scenarios where algorithms that exploit any structure in the eigenvectors of the signals, such as the MUSIC and the Capon-MVDR [5] algorithms in sensor array processing, might be better able to tease out lower level signals from the background noise. It is worth noting that the proposed approach remains relevant in situations where the eigenvector structure has been identified. This is because eigen-inference methodologies are inherently robust to eigenvector modelling errors that are endemic to high-dimensional settings. Thus the practitioner may use the proposed estimator to complement and "robustify" the inference provided by algorithms that exploit the eigenvector structure.

ACKNOWLEDGEMENTS

7. REFERENCES

[1] Mati Wax and Thomas Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 33, no. 2, pp. 387–392, 1985.

[2] I. Dumitriu and A. Edelman, "Global spectrum fluctuations for the β-Hermite and β-Laguerre ensembles via matrix models," *J. Math. Phys.*, vol. 47, no. 6, pp. 063302, 36, 2006.

[3] M. S. Bartlett, "A note on the multiplying factors for various χ^2 approximations," *J. Roy. Stat. Soc., ser. B*, vol. 16, pp. 296–298, 1954.

[4] D. N. Lawley, "Tests of significance of the latent roots of the covariance and correlation matrices," *Biometrica*, vol. 43, pp. 128–136, 1956.

[5] H. L. Van Trees, *Detection, Estimation, and Modulation Theory Part IV: Optimum Array Processing*, John wiley and Sons, Inc., new York, 2002.

[6] Eran Fishler and Hagit Messer, "On the use of order statistics for improved detection of signals by the mdl criterion," *IEEE Trans. of Signal Process.*, vol. 48, no. 8, pp. 2242–2247, August 2000.

[7] Eran Fishler and H. Vincent Poor, "Estimation of the number of sources in unbalanced arrays via information theoretic criteria," *IEEE Trans. of Signal Process.*, vol. 53, no. 9, pp. 3543–3553, September 2005.

[8] Jean-René Larocque, James P. Reilly, and William Ng, "Particle filters for tracking and unknown number of sources," *IEEE Trans. of Signal Processing*, vol. 50, no. 12, pp. 2926–2937, December 2002.

[9] N. K. Bansal and M. Bhandary, "Bayes estimation of number of signals," *Ann. Inst. Statist. Math.*, vol. 43, no. 2, pp. 227–243, 1991.

[10] Hirotugu Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, vol. AC-19, pp. 716–723, 1974, System identification and time-series analysis.

[11] George Casella and Roger L. Berger, *Statistical inference*, The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990.

[12] S. Miron, N. Le Bihan, and J.I. Mars, "Quaternion-MUSIC for vector-sensor array processing," *IEEE Trans. on Signal Process.*, vol. 54, no. 4, pp. 1218–1229, April 2006.

[13] J. Baik and J. W. Silverstein, "Eigenvalues of large sample covariance matrices of spiked population models," *Journal of Multivariate Analysis*, , no. 6, pp. 1382–1408, 2006.

[14] D. Paul, "Asymptotics of sample eigenstructure for a large dimensional spiked covariance model," Technical report, Stanford University, 2005, `http://anson.ucdavis.edu/~debashis/techrep/eigenlimit.pdf`.

[15] R. R. Nadakuditi, *Applied Stochastic Eigen-Analysis*, Ph.D. thesis, Massachusetts Institute of Technology, February 2007, Department of Electrical Engineering and Computer Science.