FACTORIZATION FOR ANALOG-TO-DIGITAL MATRIX MULTIPLICATION

Edward H. Lee[⋆]*, Madeleine Udell*[†]*, and S. Simon Wong*[⋆]

[⋆]Department of Electrical Engineering, Stanford University
[†]Institute for Computational and Mathematical Engineering, Stanford University

ABSTRACT

We present matrix factorization as an enabling technique for analog-to-digital matrix multiplication (AD-MM). We show that factorization in the analog domain increases the total precision of AD-MM in precision-limited analog multiplication, reduces the number of analog-to-digital (A/D) conversions needed for overcomplete matrices, and avoids unneeded computations in the digital domain. Finally, we present a factorization algorithm using alternating convex relaxation.

***Index Terms*—** Analog-to-digital conversion, matrix factorization, compressed sensing, analog-to-information.

1. INTRODUCTION

Analog-to-digital matrix multiplication (AD-MM) is a common task in modern sensing and communication systems. AD-MM digitizes an analog signal and multiplies the resulting data by a matrix. For example, AD-MM is used in cameras to compress digital data using transform coding and quantization [1]. Furthermore, many analog signals are known to have a sparse representation in some basis, which presents an opportunity to reduce the number of A/D conversions and the output (digital) data rate of AD-MM.

Many recent papers [2–10] have explored using analog matrix multiplication to alleviate A/D requirements for AD-MM. For example, hardware implementations of compressed sensing (CS) known as *Analog-to-Information Converters* (AIC) multiply in the analog domain a signal x, which is low-dimensional in some basis $\Psi \in \mathbb{R}^{n\times n}$, by a matrix $A \in \mathbb{R}^{m\times n}$ that is incoherent with Ψ [11, 12]. The incoherence requirement places restrictions on allowable matrices A. In practice, only Bernoulli and Hadamard matrices ($A \in \{0,1\}^{m\times n}$) are used; other allowable matrices with real-valued entries (*e.g.*, random Gaussian and partial Fourier matrices [3,13]) are usually not used due to the high precision requirements of analog hardware.

The practical limitations on the precision and size of the matrix operation not only affect AICs, but also affect other forms of analog matrix multiplication. For example, many recent papers [5–10] propose applying the discrete Fourier transform (DFT) in the analog domain for use in software defined radios. The analog DFT enables RF channelization [7] in order to relax the high RF sampling requirements from the signal-to-noise ratio (SNR) requirements, reducing the overall A/D power [14]. However, in practice, analog DFT implementations are restricted to small (*e.g.*, $n = 16$) matrices due to the large precision required.

1.1. Contributions.

In this paper, we offer factorization as a technique to increase the total precision of AD-MM while tolerating lower analog multiplication precision. We also show that analog AD-MM reduces the number of A/D conversions and can lower overall energy compared to digital (conventional) AD-MM for overcomplete matrices.

This paper is structured as follows. In §2, we present the motivations for matrix factorization. In §3, we discuss four applications of analog AD-MM using factorization. In §4, we present an algorithm to factor large matrices and show that it can efficiently compute a good factorization requiring only modest analog precision.

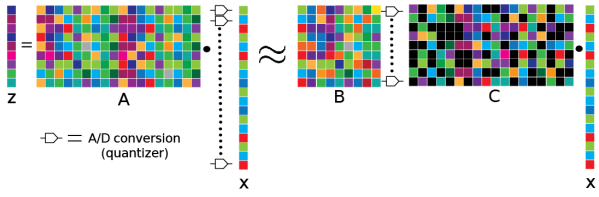

Fig. 1. Digital AD-MM (left) and analog AD-MM (right).

1.2. Definitions and notation.

Formally, digital (conventional) AD-MM as shown in Fig. 1 applies a matrix A to a digitized representation of the analog signal x. The digitization process is described by the quantization function Q. Analog AD-MM, proposed in this paper, instead relies on a factorization $BC \approx A$ with $B \in \mathbb{R}^{m\times m}$ and $C \in \mathbb{R}^{m\times n}$. Analog AD-MM applies C in the analog

This work was supported by the Texas Instruments Stanford Graduate Fellowship, the Gerald J. Lieberman Fellowship, the Gabilan Stanford Graduate Fellowship, and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

domain to the analog data $x \in \mathbb{R}^n$, quantizes the result, and finally applies B in the digital domain to produce the output z. Our goal is to design B and C such that $BQ(Cx) \approx AQ(x)$. Throughout this paper, we restrict our attention to square matrices B, although extensions to rectangular B are possible, and usually consider the overcomplete case $m < n$.

2. MATRIX FACTORIZATION: MOTIVATIONS

In digital matrix multiplication, an increase in precision (or resolution) necessitates an increase in the number of digital control gates, size of memory, and interconnects, all of which increase the energy of a multiply-and-accumulate (MAC) operation [15]. Energy also increases with precision in analog matrix multiplication. A prototype analog multiplier (Fig. 2) illustrates how each voltage x_j is sampled on the capacitor array connected to the signal path at time j to form a charge proportional to the total capacitance of the array. We let c_{ij} be the constant of proportionality. This sampling operation is performed for each $j = 1 : n$ and accumulated together in the analog charge domain to form $y_i = \Sigma_{j=1}^{n} c_{ij} x_j$. Analog charge domain multiplication is a practical approach for analog MAC, although many other approaches and variants are possible [10, 16]. Nonetheless, with any of these approaches, an increase in precision of the analog matrix C requires an increase in the number of different analog multiplier elements (*e.g.*, capacitors in Fig. 2), for the roughly the same reasons as in the digital case. Therefore, energy grows with precision just as it does in digital domain matrix multiplication.

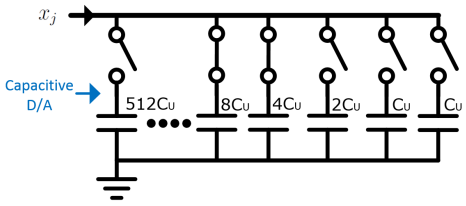

Fig. 2. Example of an analog charge-domain multiplier with binary-weighted encoding ($c_{ij} = \frac{12}{1024}$ shown) with precision $N_C = 10$ bits (2^{10} different multiplier values).

Factorization can relax the hardware requirements imposed by a need for high precision. Formally, to find a good factorization, we solve for B and C in

$$\begin{array}{ll} \text{minimize} & ||A - BC||_F \\ \text{subject to} & C \in \Omega_C \\ & B \in \Omega_B, \end{array} \tag{1}$$

where C (B) is constrained to take on values in an integer set Ω_C (Ω_B). For example, Fig. 2 gives an example using binary-weighted encoding, *i.e.*,

$$c_{ij} = \tfrac{c_{max}}{2^{N_C}} y_{ij}, \quad -2^{N_C} \le y_{ij} \le 2^{N_C}, \quad y_{ij} \in \mathbb{Z},$$

for $i = 1, ..., n$ and $j = 1, ..., m$, where N_C is the multiplier precision. The constraint sets (Ω_B, Ω_C) can also enforce non-negativity and other constraints to capture the hardware limitations of analog multiplication.

How much precision do we require in C? The answer depends on the application — for example, a coarse multiplication where $C = Q(A)$ at the 2 bit level is good enough for detection (see §3) while 10 bits may be required to meet stringent SNR requirements.

Recall that the dimensionality of the digital output space is size m, while the dimensionality of the analog input space is size n, where $m < n$. Thus analog AD-MM requires only m A/D conversions, a significant savings over the n A/D conversions required by a digital AD-MM system. Analog AD-MM also requires a digital computation (*i.e.*, multiplication by B) of size m^2, instead of nm for digital AD-MM, but has the added overhead of the analog multiplication by C of size nm.

Furthermore, *pruning* (eliminating outputs $(Cx)_j \approx 0$) also reduces the number of digital MACs required. If we can detect $(Cx)_j \approx 0$, then we need not multiply it by the j-th row of B. For example, if $|(Cx)_j| \le \delta$ for all $j = 1, \ldots, m$ and δ is an application-specific pruning threshold, then multiplication by B need not be performed (see §3).

Thus, compared to digital AD-MM, analog AD-MM requires fewer A/D conversions and fewer digital MACs, but incurs the extra overhead of the analog multiplication Cx. To first order, the circuit-independent energy estimates for digital (E_{d}) and analog (E_{a}) AD-MM are

$$\begin{aligned} E_{\text{d}} &\simeq mnE_{\text{d-op}} + nE_{\text{A/D}} \\ E_{\text{a}} &\simeq mnE_{\text{a-op}} + mE_{\text{A/D}} + (\gamma m)^2 E_{\text{d-op}}, \end{aligned}$$

where $E_{\text{d-op}}$ is the energy cost per digital MAC operation (Joule/op), $E_{\text{A/D}}$ is the cost per A/D conversion, $E_{\text{a-op}}$ is the cost per analog MAC operation, and $\gamma \in [0, 1]$ is the pruning factor. We assume here that the A/D quantizes the analog signal to the same precision for both architectures. (All three energy costs depend strongly on the precision.)

The first-order estimates show that increasing n increases both the digital MAC and the A/D costs in digital AD-MM but increases only the analog MAC cost in analog AD-MM. Furthermore, [7, 8] report that $E_{\text{a-op}} < E_{\text{d-op}}$, which makes analog AD-MM attractive from an energy perspective as well. For example, the charge-domain FFT proposed in [7] achieves an energy per analog FFT operation that is 130 times lower than its digital equivalent, at similar SNR levels.

3. EXAMPLES

Using four examples, we show that lowering the precision in B and C need not decrease the fidelity of the output z.

3.1. Factorization for precision-limited C

As a first example (Fig. 3), we factor a truncated discrete cosine transform (DCT) matrix $A \in \mathbb{R}^{50\times120}$, keeping the 50 basis vectors (rows) of lowest frequency, and constrain the factor C to have binary-weighted elements with precisions $N_C = 10$, $N_C = 6$, and $N_C = 4$ bits. The factorizations decrease in quality at lower precision, achieving MSEs of 4.8×10^{-5}, 6.6×10^{-4}, and 1.3×10^{-3} for $N_C = 10$, 6, and 4 respectively.

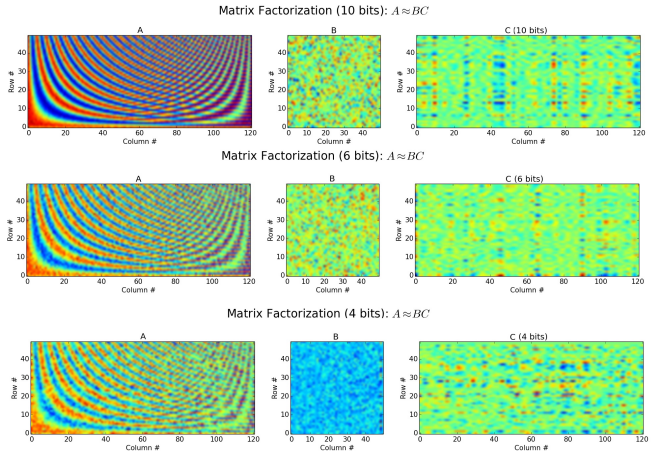

Fig. 3. Factorization results on the truncated DCT matrix.

3.2. Image Reconstruction

As a second example (Fig. 4), we factor a truncated DCT matrix $A \in \mathbb{R}^{16\times64}$ with $N_C = 10$ bit for an application in image reconstruction. The resulting matrix BC is used to reconstruct the image for each color channel (RGB) independently. The reconstruction shows little loss in fidelity, despite a 4:1 compression ratio. The peak SNR decreases from 28.3 dB (original) to 26.7 dB (reconstruction).

Fig. 4. Reconstruction of an image on 8×8 patches.

3.3. Factorization for precision-limited B and C

As our third example, we show that factorization allows us to use *lower* precision in our intermediate MAC operations while maintaining approximately the same output error in z. In Fig. 5, we examine the effect of bounded precision in A, B and C on fidelity of the output, using fixed-point (FP) arithmetic. The data $A \in \mathbb{R}^{20\times200}$ is generated by selecting entries uniformly from $[-1, 1]$. The desired output is $z_{\text{true}} = Ax$ where x is a 10-bit FP i.i.d. Gaussian random vector. To factor A into B and C, we solve Prob. 1 with the constraint that $||B||_\infty \leq 1$. Define $z_A = Q_{N_A}(A)x$ and $z_{BC} = Q_{N_B}(B)Q_{N_C}(C)x$, where Q_N denotes quantization to N bits in FP, and all MAC operations are performed in FP arithmetic. We measure output error on output z as the average of the loss $L(z) = ||z - z_{\text{true}}||_2$ over 1000 realizations of the input x.

Fig. 5 shows $\mathbb{E}[L(z_A)]$ and $\mathbb{E}[L(z_{BC})]$ as we vary the precision N_B of B while fixing the precision of A, C, and Cx. We achieve similar expected loss using factored AD-MM with $N_B = 5$ and $N_C = 4$ or using digital AD-MM with $N_A = 5$. For higher N_B, FP factored AD-MM approaches the performance of infinite precision factored AD-MM. We can further decrease the expected loss by finding a better factorization of A into low precision factors; we return to this question in §4.

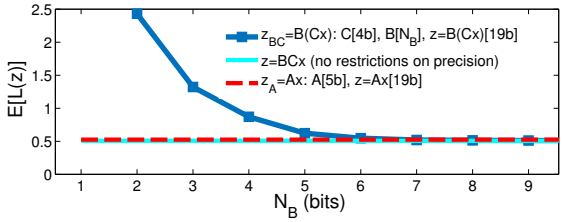

Fig. 5. Expected output error for FP factored AD-MM and FP digital AD-MM. Here, $A[N_A]$ denotes A quantized to precision N_A in FP.

3.4. Detection and estimation

Many sensing applications (*e.g.*, radar) require the ability to actively detect a signal's presence and the ability to estimate signal features. If the detection accuracy is robust to decreased precision, low precision sensing can save energy while reliably detecting the signal. However, high precision may be necessary to estimate signal features once the signal is detected.

As an example of a detection-estimation problem, consider a generalized likelihood ratio test (GLRT) [17]. The task is to classify whether the signal is present or not, and to estimate the (unknown) time of arrival ($s_{\Delta t}$). Let the observation vector be $x \in R^n$, the signal be $s \in R^n$, and the noise be $w \sim \mathcal{N}(0, \sigma^2 I)$ (Fig. 6 (b)). The GLRT (with threshold η) rejects the null hypothesis $\mathcal{H}_0$ (no signal) in favor of $\mathcal{H}_1$ (signal) if $\frac{p(x;\Delta\hat{t},\mathcal{H}_1)}{p(x;\mathcal{H}_0)} > \eta$ for all $\Delta\hat{t}$. The maximum likelihood estimate of the time of arrival (ToA) is $\Delta\hat{t} = \arg\max_j a_j^T x$, where a_j^T is a row vector containing a time-shifted version of s. These row vectors are collected in the matrix A (see Fig. 6 (a)). For detection, we let C be the positive part of $Q_{N_C}(A)$, *i.e.*, the matrix A quantized to N_C bits. We force C to be nonnegative,

since an analog multiplier in practice generally requires additional control logic to encode negative values. Then, for ToA estimation, we set $B = \text{argmin}_{B\in\Omega_B} ||A - BC||_F$, where $\Omega_B = \mathbb{R}^{m\times m}$.

We show in Fig. 6(c) only a modest decrease in detection accuracy for a 1-bit, nonnegative C at a given SNR compared to using full precision A. Furthermore, using the backend matrix B significantly increases the time of arrival (ToA) estimation accuracy, conditioned on detection of the signal (see Fig. 6(d)). For example, with an SNR $= -7.8$ dB, we see an increase from 74% to 91% ToA estimation. Thus, for this particular application of AD-MM, reducing precision is practical since it lowers the analog AD-MM complexity while preserving detection performance.

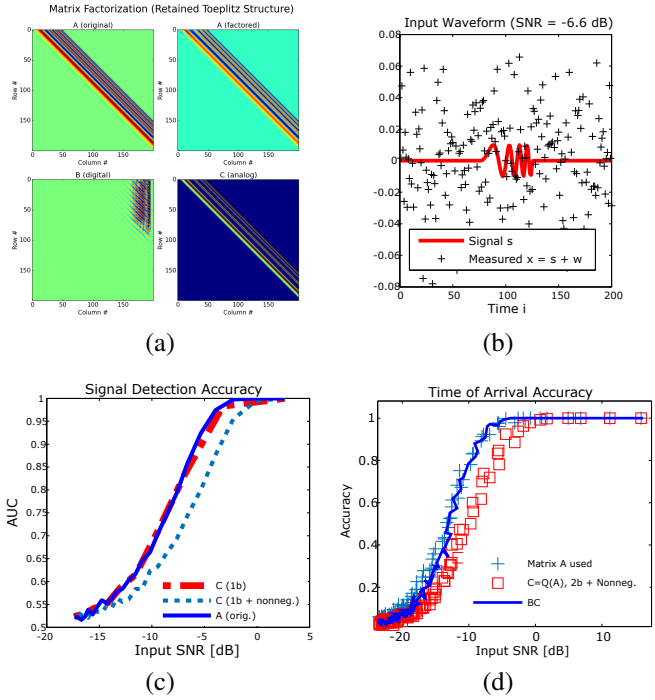

Fig. 6. (a) Factorization for A where C is constrained to be nonnegative, with 2 bit precision. (b) The input waveform. (c) Area Under the Curve (AUC) found from Receiver Operating Characteristic curves. (d) ToA accuracy.

4. COMPUTING THE FACTORIZATION

For most applications with stationary signal statistics, factorization need only be performed once, but must be done well. Unfortunately, the factorization problem in Eq. 1 is not convex, due to 1) the non-convex constraints $C \in \Omega_C$ and $B \in \Omega_B$, and 2) the product of variables BC. To find an approximate solution for one variable holding the other fixed, we use a relax-and-round heuristic: minimize over the convex hull of the feasible set (*i.e.*, $\mathbf{conv}(\Omega)$), and quantize the resulting matrix with N bit precision. We then use alternating minimization to find an approximate solution for the full problem:

1: **repeat**
2: $\quad B^{(k)} \leftarrow Q_{N_B}\left(\text{argmin}_{B\in\mathbf{conv}(\Omega_B)} ||A - BC||_F\right)$
3: $\quad C^{(k)} \leftarrow Q_{N_C}\left(\text{argmin}_{C\in\mathbf{conv}(\Omega_C)} ||A - BC||_F\right)$
4: $\quad \epsilon_{k+1} \leftarrow ||A - B^{(k)}C^{(k)}||_F$
5: **until** converged

We compute the approximation error $\epsilon = ||A-BC||_F$ for factorizations of 50 randomly initialized (RI) matrices, where C is constrained to lie in a nonnegative, FP set. The final distribution of ϵ for different precisions shown in Fig. 7(a), and indicates that ϵ converges to a compact minimum at 10 bit precision after just a few iterations. However, for smaller precisions, the average ϵ is significantly worse (see Fig. 7(b)).

Greedy initialization (GI) overcomes this problem. GI uses the locally optimal matrices $C^{(k-1)}$ found at higher precisions $N_C + 1$ and $N_B + 1$ to initialize C in the alternating minimization to initialize the search for a new factorization with precisions N_C and N_B. Fig. 7(b) shows that the GI performs substantially better than RI at low precision.

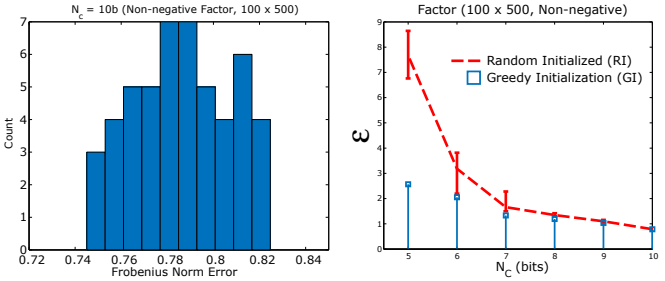

Fig. 7. (a) Histogram using the RI factorization procedure. (b) ϵ for different precisions N_C using RI and GI.

5. CONCLUSION

Factorization is an enabling technique for analog AD-MM that increases its advantages over (conventional) digital AD-MM. Factorization can increase the total precision of analog AD-MM even with lower analog multiplication precision. Examples show that analog AD-MM performs well and that a good factorization requiring only modest analog precision can be efficiently computed. The authors are currently working to implement analog AD-MM in hardware.

6. ACKNOWLEDGEMENTS

We thank Boris Murmann, Chris Young, Doug Adams, Niki Hammler, Daniel Bankman, Thomas Lipp, Stephen Boyd, and Abbas El Gamal for their helpful comments.

7. REFERENCES

[1] A. El Gamal and H. Eltoukhy, "CMOS image sensors," *IEEE Circuits and Devices Magazine*, vol. 21, no. 3, pp. 6–20, May 2005.

[2] D. Adams, C. S. Park, Y. C. Eldar, and B. Murmann, "Towards an integrated circuit design of a Compressed Sampling wireless receiver," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 5305–5308.

[3] M. Herman and T. Strohmer, "Compressed sensing radar," in *IEEE Radar Conference, 2008. RADAR '08.*, May 2008, pp. 1–6.

[4] Y. Oike and A. El Gamal, "CMOS Image Sensor with Per-Column $\Sigma\Delta$ ADC and Programmable Compressed Sensing," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 318–328, Jan 2013.

[5] A. A. Abidi, "The Path to the Software-Defined Radio Receiver," *IEEE Journal of Solid-State Circuits*, vol. 42, no. 5, pp. 954–966, May 2007.

[6] B. Sadhu, M. Sturm, B. M. Sadler, and R. Harjani, "A 5GS/s 12.2pJ/conv. analog charge-domain FFT for a software defined radio receiver front-end in 65nm CMOS," in *2012 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, June 2012, pp. 39–42.

[7] B. Sadhu, M. Sturm, B. M. Sadler, and R. Harjani, "Analysis and Design of a 5 GS/s Analog Charge-Domain FFT for an SDR Front-End in 65 nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 5, pp. 1199–1211, May 2013.

[8] Y-W. Lin, H-Y. Liu, and C-Y. Lee, "A 1-GS/s FFT/IFFT processor for UWB applications," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 8, pp. 1726–1735, Aug 2005.

[9] M. Lehne and S. Raman, "A 0.13μm 1-GS/s CMOS Discrete-Time FFT processor for Ultra-Wideband OFDM Wireless Receivers," *IEEE Transactions on Microwave Theory and Techniques*, vol. 59, no. 6, pp. 1639–1650, June 2011.

[10] F. Rivet, Y. Deval, J. Begueret, D. Dallet, P. Cathelin, and D. Belot, "The Experimental Demonstration of a SASP-Based Full Software Radio Receiver," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 5, pp. 979–988, May 2010.

[11] S. Kirolos, J. Laska, M. Wakin, M. Duarte, D. Baron, T. Ragheb, Y. Massoud, and R. Baraniuk, "Analog-to-Information Conversion via Random Demodulation," in *2006 IEEE Dallas/CAS Workshop on Design, Applications, Integration and Software*. IEEE, 2006, pp. 71–74.

[12] O. Abari, F. Lim, F. Chen, and V. Stojanovic, "Why Analog-to-Information Converters suffer in high-bandwidth sparse signal applications," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 9, pp. 2273–2284, 2013.

[13] W. Yin, S. Morgan, J. Yang, and Y. Zhang, "Practical compressive sensing with Toeplitz and circulant matrices," in *Visual Communications and Image Processing 2010*. International Society for Optics and Photonics, 2010, pp. 77440K–77440K.

[14] B. Murmann, "ADC Performance Survey 1997-2014," [Online]. Available: `http://web.stanford.edu/~murmann/adcsurvey.html`.

[15] S. Galal and M. Horowitz, "Energy-Efficient Floating-Point Unit Design," *IEEE Transactions on Computers*, vol. 60, no. 7, pp. 913–922, July 2011.

[16] W. Xiong, U. Zschieschang, H. Klauk, and B. Murmann, "A 3V 6b successive-approximation ADC using complementary organic thin-film transistors on glass," in *2010 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, Feb 2010, pp. 134–135.

[17] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.