

Contents

<i>Preface</i>	page xv
<i>Acknowledgments</i>	xix
1 Getting Started	1
1.1 Introduction	1
1.2 Example Applications	1
1.2.1 Signal Processing Example: Image Deblurring	1
1.2.2 Computer Vision Applications	2
1.2.3 Machine Learning Example: Handwritten Digit Recognition	3
1.3 Formatting	3
1.4 Notation Preview	4
1.4.1 What the $\mathbb{F}$ Means	6
1.5 Julia	7
1.6 Fields, Vector Spaces, Linear Maps	7
1.6.1 Field of Scalars	8
1.6.2 Vector Spaces	9
1.6.3 Linear Maps and Linear Operators	11
2 Introduction to Matrices	12
2.1 Introduction	12
2.2 Basics of Vectors and Matrices	12
2.3 Matrix Structures	17
2.3.1 Common Matrix Shapes and Types	17
2.3.2 Matrix Transpose and Symmetry	20
2.4 Multiplication	23
2.4.1 Vector–Vector Multiplication	23
2.4.2 Matrix–Vector Multiplication	25
2.4.3 Matrix–Matrix Multiplication	29
2.4.4 Matrix Multiplication Properties	30
2.4.5 Kronecker and Hadamard Products, and the vec Operator	34
2.4.6 Using Matrix–Vector Operations	35
2.4.7 Invertibility	40
2.5 Orthogonality and the Euclidean Norm	42
2.5.1 Orthogonal Vectors	42
2.5.2 Euclidean Norm	42

2.5.3	Cauchy–Schwarz Inequality	43
2.5.4	Orthogonal Matrices	43
2.6	Determinant of a Matrix	45
2.6.1	Determinant Properties	46
2.6.2	Matrices with Units	47
2.6.3	Laplace’s Determinant Formula	48
2.6.4	Avoiding Computation	48
2.6.5	Small Matrices	49
2.7	Eigenvalues	50
2.7.1	Eigenvectors	51
2.7.2	Practical Implementation	51
2.7.3	Properties of Eigenvalues	52
2.8	Trace	54
2.9	Summary	55
3	Matrix Factorization: Eigendecomposition and SVD	63
3.1	Introduction	63
3.1.1	Matrix Factorizations	63
3.1.2	Square Matrices	64
3.2	Spectral Theorem for Symmetric Matrices	65
3.2.1	Normal Matrices	66
3.2.2	Square Asymmetric and Nonnormal Matrices	68
3.2.3	Geometry of Matrix Diagonalization	70
3.2.4	Matrix Powers and Matrix Exponential	73
3.3	Singular Value Decomposition	74
3.3.1	Singular Values and Singular Vectors	74
3.3.2	Existence of SVD	74
3.3.3	Geometry	75
3.3.4	Practical Implementation in JULIA	76
3.3.5	SVD Basic Properties	77
3.4	The Matrix 2-Norm or Spectral Norm	78
3.4.1	Optimization: min versus arg min	79
3.4.2	Eigenvalues as Optimization Problems	80
3.4.3	Smallest Singular Value	81
3.5	Relating SVDs and Eigendecompositions	82
3.5.1	When Does $U = V$?	84
3.5.2	SVD Computation Using Eigendecomposition	85
3.5.3	SVD Nonuniqueness Revisited	87
3.6	Positive Semidefinite Matrices	88
3.6.1	Relating Positive (Semi)Definiteness to Eigenvalues	89
3.7	Summary	89
4	Subspaces, Rank, and Nearest-Subspace Classification	96
4.1	Introduction	96
4.2	Subspaces	96
4.2.1	Span	98

4.2.2	Linear Independence	99
4.2.3	Basis	101
4.2.4	Dimension	103
4.2.5	Sums and Intersections of Subspaces	105
4.2.6	Direct Sum of Subspaces	106
4.2.7	Dimensions of Sums of Subspaces	106
4.2.8	Orthogonal Complement of a Subspace	107
4.2.9	Linear Transforms	108
4.2.10	Range of a Matrix	109
4.3	Rank of a Matrix	111
4.3.1	Practical Use in JULIA	112
4.3.2	Rank of a Matrix Product	112
4.3.3	Other Rank Properties	113
4.3.4	Spark	114
4.3.5	Unitary Invariance of Rank	114
4.4	Nullspace of a Matrix	115
4.4.1	Nullspace or Kernel	116
4.4.2	Properties of Null Space	116
4.4.3	Columns of Unitary Matrices	118
4.5	The Four Fundamental Spaces	119
4.5.1	Anatomy of the SVD	121
4.5.2	SVD of Finite Differences	123
4.5.3	Synthesis View of Matrix Decomposition	125
4.6	Orthogonal Bases	125
4.6.1	Finding Coordinates in an Orthogonal Basis	126
4.6.2	Stiefel Manifold of Orthogonal Bases	127
4.7	Spotting Decompositions	128
4.7.1	Matrix–Vector Products and the SVD	130
4.8	Application: Signal Classification	130
4.8.1	Projection onto a Set	130
4.8.2	Nearest Point in a Subspace	131
4.8.3	Signal Classification by Nearest Subspace	133
4.9	Optimization Preview	134
4.9.1	Convex Sets	135
4.9.2	Convex Functions	136
4.10	Summary	137
5	Linear Least-Squares Regression and Binary Classification	143
5.1	Introduction	143
5.2	Introduction to Linear Equations	143
5.2.1	Solving $Ax = y$	144
5.2.2	Linear Regression and Machine Learning	144
5.2.3	Lifting for Nonlinear Regression	145
5.3	Linear Least-Squares Estimation	145
5.3.1	Minimization and Gradients	148
5.3.2	Solving LLS Using the Normal Equations	150

5.3.3	Solving LLS Problems Using the Compact SVD	151
5.3.4	Uniqueness of LLS Solution	154
5.4	Moore–Penrose Pseudoinverse	155
5.4.1	Pseudoinverse and Matrix Products	155
5.4.2	Pseudoinverse and SVD	156
5.5	LLS: Under-Determined Case	158
5.5.1	Orthogonality Principle	160
5.5.2	Minimum-Norm LS Solution via Pseudoinverse	162
5.6	Truncated SVD Solution	164
5.6.1	Condition Number	164
5.6.2	Practical Implementation of Truncated SVD Solution	165
5.6.3	Low-Rank Approximation Interpretation of Truncated SVD	165
5.6.4	Noise Effects and Perturbations	166
5.6.5	Tikhonov Regularization, or Ridge Regression	167
5.7	Summary of LLS Solution Methods in Terms of SVD	168
5.8	Frames and Tight Frames	168
5.8.1	Properties of a Frame	170
5.8.2	Tight Frame	170
5.8.3	Parseval Tight Frame	171
5.8.4	Properties of Parseval Tight Frames	171
5.8.5	Frame Summary	173
5.9	Projection and Orthogonal Projection	174
5.9.1	Idempotent Matrix	174
5.9.2	Orthogonal Projection Matrix	176
5.9.3	Projection onto a Subspace	177
5.9.4	Binary Classifier Design Using Least Squares	182
5.9.5	Empirical Risk Minimization	183
5.10	Recursive Least Squares	184
5.10.1	RLS with a Forgetting Factor	185
5.11	Summary	186
6	Norms and Procrustes Problems	197
6.1	Introduction	197
6.2	Vector Norms	197
6.2.1	Examples of Vector Norms	198
6.2.2	Practical Implementation	199
6.2.3	Properties of Vector Norms	200
6.2.4	Norm Notation	201
6.2.5	Robust Regression Application	201
6.2.6	Unitarily Invariant Vector Norms	202
6.3	Inner Products	203
6.3.1	Examples of Inner Products	203
6.3.2	Properties of Inner Products	204
6.3.3	More Inner Product Inequalities	205
6.3.4	Angle Between Vectors	205
6.3.5	Angle Between Subspaces	206

6.4	Matrix Norms and Operator Norms	206
6.4.1	Examples of Matrix Norms	207
6.4.2	Induced Matrix Norms	208
6.4.3	Norms Defined in Terms of Singular Values	210
6.4.4	Practical Implementation	214
6.4.5	Properties of Matrix Norms	214
6.4.6	Spectral Radius	217
6.4.7	Practical Step Size for Gradient Descent	218
6.5	Convergence of Sequences of Vectors and Matrices	219
6.6	Generalized Inverse of a Matrix	221
6.6.1	Minimum Frobenius Norm Generalized Inverse	221
6.7	Procrustes Analysis	222
6.7.1	Sanity Check and Scale Invariance	224
6.7.2	Procrustes Generalizations	225
6.7.3	Subspace/Span Comparisons	228
6.7.4	Weighted Procrustes Problems	228
6.7.5	Practical Implementation	229
6.8	Summary	229
7	Low-Rank Approximation and Multidimensional Scaling	238
7.1	Introduction	238
7.2	Low-Rank Approximation via Frobenius Norm	238
7.2.1	Eckart–Young–Mirsky Theorem	239
7.2.2	Subspace Approximation Perspective	240
7.2.3	Implementation	240
7.2.4	Choosing Rank via Permutation	242
7.2.5	Nonuniqueness of SVD and Low-Rank Approximation	243
7.2.6	One-Dimensional Example	244
7.2.7	Generalization to Other Norms	245
7.2.8	Bases for $\mathbb{F}^{M \times N}$	247
7.2.9	Low-Rank Approximation Summary	248
7.2.10	Rank and Stability	249
7.2.11	Example: Photometric Stereo	250
7.3	Sensor Localization Application: Multidimensional Scaling	250
7.3.1	Derivation (Analysis)	251
7.3.2	MDS Method	254
7.3.3	Practical Implementation	255
7.3.4	Extensions	256
7.4	Proximal Operators	257
7.4.1	Soft Thresholding	257
7.4.2	Hard Thresholding	259
7.5	Alternative Low-Rank Approximation Formulations	260
7.5.1	Unconstrained/Regularized Formulation	260
7.5.2	General Unitarily Invariant Formulations	260
7.5.3	Singular Value Hard Thresholding	261
7.5.4	Singular Value Soft Thresholding	262

7.5.5	Other Extensions of Low-Rank Approximation	263
7.6	Choosing the Rank or Regularization Parameter	263
7.6.1	Stein's Unbiased Risk Estimate	264
7.6.2	OptShrink	265
7.7	Related Methods: Autoencoders and PCA	269
7.7.1	Relation to Autoencoder with Linear Layers	269
7.7.2	Relation to Principal Component Analysis	270
7.8	Subspace Learning for Classification	275
7.8.1	Subspace Clustering	277
7.9	Subspace Tracking and Streaming PCA	277
7.9.1	Incremental SVD	278
7.9.2	Streaming PCA	278
7.10	Summary	279
8	Special Matrices, Markov Chains, and PageRank	283
8.1	Introduction	283
8.2	Companion Matrices	283
8.2.1	Practical Implementation	285
8.2.2	Polynomial Matrix Functions	286
8.2.3	Eigenvectors of Companion Matrices	287
8.2.4	Vandermonde Matrices	288
8.2.5	Kronecker Sum and Polynomial Roots	289
8.3	Circulant Matrices	290
8.3.1	Relationship to DFT Properties from DSP	293
8.3.2	Practical Implementation	293
8.3.3	Spectral Properties of Circulant Matrices	294
8.3.4	Inverting a Circulant Matrix	294
8.4	Toeplitz Matrices	294
8.4.1	Toeplitz Matrix Multiplication with a Vector	295
8.4.2	Inverting a Toeplitz Matrix	295
8.4.3	Factoring a Toeplitz Matrix	295
8.5	Power Iteration	296
8.5.1	Convergence of the Power Iteration	297
8.5.2	Geršgorin Disk Theorem	298
8.6	Nonnegative Matrices and Graphs	301
8.6.1	Primitive Matrices	301
8.6.2	Weighted Directed Graphs	303
8.6.3	Strongly Connected Graphs	305
8.6.4	Irreducible Matrix	306
8.6.5	Matrix Period	307
8.7	Nonnegative Matrices and Perron–Frobenius Theorems	309
8.7.1	Perron–Frobenius for Square Nonnegative Matrices	309
8.7.2	Perron–Frobenius for Nonnegative Irreducible Matrices	311
8.7.3	Perron–Frobenius for Primitive Matrices	312
8.7.4	Perron–Frobenius for Stochastic Matrices	312
8.8	Markov Chains	313

8.8.1	Equilibrium Distribution(s) of a Markov Chain	315
8.8.2	Limiting Distribution(s) of a Markov Chain	316
8.8.3	Markov Chains with Strongly Connected Graphs	318
8.8.4	Google's PageRank Method	319
8.9	Graph Laplacian and Spectral Clustering	322
8.9.1	Clustering	322
8.9.2	Weighted Graph Based on Similarity	323
8.9.3	Connected Components	324
8.9.4	Graph Laplacian	324
8.9.5	Spectral Clustering Algorithm	325
8.9.6	Laplacian Eigenmaps	326
8.10	Summary	327
9	Optimization Basics and Logistic Regression	335
9.1	Introduction	335
9.2	Preconditioned Gradient Descent for LS	335
9.2.1	Tool: Matrix Square Root	336
9.2.2	Convergence Rate Analysis of PGD: First Steps	338
9.2.3	Classical GD: Step Size Bounds	339
9.2.4	Optimal Step Size for GD	340
9.2.5	Ideal Preconditioner for PGD	340
9.2.6	Tool: Positive (Semi)Definiteness Properties	341
9.2.7	General Preconditioners for PGD	342
9.2.8	Diagonal Majorizer	342
9.2.9	Preconditioning Illustration/Demo	344
9.2.10	Convergence Rates	345
9.2.11	Tool: Commuting (Square) Matrices	346
9.2.12	Monotonicity	347
9.3	Preconditioned Steepest Descent	349
9.4	Gradient Descent for Smooth Convex Functions	349
9.4.1	Lipschitz Continuity	350
9.4.2	Convexity and Hessian	351
9.4.3	GD Convergence Theorem	352
9.4.4	Nesterov's Fast Gradient Method	353
9.4.5	Optimized Gradient Method	354
9.4.6	Gradient Projection Method	355
9.5	Machine Learning via Logistic Regression for Binary Classification	356
9.5.1	Practical Implementation of Logistic Regression	360
9.5.2	Numerical Results: Logistic Regression	360
9.6	Stochastic Gradient Descent	360
9.7	Summary	361
10	Matrix Completion and Recommender Systems	365
10.1	Introduction	365
10.2	Measurement Model	366
10.2.1	Practical Implementation	366
10.2.2	Sampling Conditions for LRMC	366

10.2.3	Sampling Mask	367
10.3	LRMC: Noiseless Case	368
10.3.1	Noiseless Problem Statement	368
10.3.2	Alternating Projection Approach to LRMC	368
10.4	LRMC: Noisy Case	371
10.4.1	Noisy Problem Statement	371
10.4.2	Majorize–Minimize (MM) Iterations	372
10.4.3	MM Methods for LRMC	373
10.4.4	LRMC by Iterative Low-Rank Approximation	373
10.4.5	LRMC by Iterative Singular Value Hard Thresholding	374
10.4.6	LRMC by Iterative Singular Value Soft Thresholding	374
10.4.7	Iterative Soft-Thresholding Algorithm	375
10.4.8	Debiasing the Nuclear Norm Effects	376
10.4.9	Factorization Approaches	377
10.4.10	Demo	378
10.5	Robust PCA and Video Foreground/Background Separation	378
10.5.1	Robust PCA	378
10.5.2	Video Foreground/Background Separation	378
10.6	Nonnegative Matrix Factorization	379
10.7	Summary	380
11	Neural Network Models	381
11.1	Introduction	381
11.2	The Importance of Nonlinearity	381
11.3	Fully Connected NN Models	383
11.3.1	Perceptron Model	383
11.3.2	Multilayer Perceptron NN Models	384
11.3.3	Model Expressiveness	385
11.4	Training NN Models	385
11.4.1	Weight Regularization	386
11.5	CNN Models	387
11.5.1	Matrix Representations	388
11.5.2	CNN Architectures	388
11.6	Summary	389
12	Random Matrix Theory, Signal + Noise Matrices, and Phase Transitions	390
12.1	Introduction	390
12.1.1	Perturbation Bounds	390
12.2	Roundoff Error	391
12.2.1	RMT for Roundoff Analysis	393
12.3	Additive Noise	396
12.4	Outliers	400
12.5	Matrix Completion	401
12.6	Summary	403
	<i>References</i>	405
	<i>Index</i>	423