

Index

- A-optimality, *see* optimal design
- a.s.w., *see* average silhouette width
- Absolute correlation, 161
- Active genes, 177
- Adaptive circle segmentation, 61
- Adaptive shape segmentation, 61
- Addressing, *see* gridding
- Adenine, 15
- Affected genes, 177
- affy, 95, 101
- Affymetrix
 - data normalization, 94–101
 - data structure, 93–94
 - Gene Chips, 10, 16, 24, 28, 40, 93–101
 - MAS 5.0, 10, 43, 94, 96, 98, 100
 - reading data into R, 94
- Agglomerative hierarchical
 - clustering, 146–148, 163–165
- AIC, *see* Akaike information criterion
- Akaike information criterion (AIC), 172
- all.norm, 70
- Alternative hypothesis, 178, 180
- (Analysis of variance), *see* ANOVA
- ANOVA (analysis of variance), 63
- Arabidopsis, 16
- Arabidopsis experiment, 4–6, 16–21
- Arabidopsis thaliana*, *see* Arabidopsis
- Array effect normalization
 - across-condition, 88–93
 - global, 85–86, 93, 98, 99
 - within-replication, 86–88, 92–93
- Artifacts
 - detection, 66, 103, 121–122
 - high-end, 64, 121
 - low-end, 64, 121
- Average silhouette width (a.s.w.), 152, 167
- Background contamination, 77
- Background correction
 - additive, 76
 - dangers of, 76–77
 - filtering, 59
 - global, 77–80, 97
 - local, 76–77
 - one channel, 96–97
 - two-channel, 76–80
- Background subtraction, 76
- ‘Banana effect’, 70, 80
- Bayes classifiers, 212
- Bayesian FDR procedure, 199
- Bayesian hypothesis testing, *see* hypothesis testing
- Bayesian inference, 178
- Bayesian information criterion (BIC), 172
- Benjamini & Hochberg FDR procedure, 195, 201–202
- Bias, 24, 30

- BIC, *see* Bayesian information criterion
- Bioconductor, 2, 79
- bkg.norm, 79–80, 97
- Bleeding, *see* carry-over effect
- Blocking factor, 37–38
- Bonferroni correction, 194
 - myths, 247–248
- Bootstrap, 185
- Bootstrap *p*-values, 184–186
- Breast cancer experiment, 7–9
- Breast tissue
 - tumours, *see* breast cancer experiment
- CAGED algorithm, 163, 174–175
- Cancer
 - breast, *see* breast cancer experiment
 - skin, *see* skin cancer experiment
- Carry-over effect (bleeding), 40, 120
- CDF files (.cdf), 93–94
- cDNA (complementary deoxyribonucleic acid), 18
 - folding, 20
 - labelled, 19
- CEL files (.cel), 93–94
- cel2int, 95
- Centroid, 167
- CLARA, 168
- Classification, 211
 - Bayes rule, 213
 - continuous, 217–218
 - probabilistic, 212–217
- Clerical errors, 114–117, 123
- cluster, 147
- Cluster validation, 152–154
- cluster.samples, 150
- Clustering
 - dissimilarity measures
 - samples, 139–144
 - genes, 160–175
 - model-based, 171–175
 - samples, 137–155
 - methods, 144–155
 - two-way, 165–167
 - visualization, 144
- Co-expression, 161
- Co-regulation, 160
- Complementary deoxyribonucleic acid, *see* cDNA
- Composite sequence, 104
- cond.norm, 92–93, 99
- Condition, 37, 41
- Confounding
 - design, 37–38
 - spatial and dye effect, 132–133
- Constant background subtraction,
 - see* background correction. global
- Control spots, 65
- Correlation, *see* dissimilarity measures *and* co-expression
- Correlograms, 113
- Cosine distance, 161
- Cross-validation
 - generalized, 239
 - K*-fold, 213, 216–217, 238–240
 - leave-one-out, 214, 217, 239
 - myths, 249
- Crossing factor, 37–38
- cv.lars, 241–243
- Cy3/Cy5 dyes, 19
- Cytosine, 15
- D-optimality, *see* optimal design
- Data augmentation, 68
- dChip, 66, 94, 103
- Decision rules
 - Bayesian, 199–200
 - for multiple genes, 193–195, 199
 - for single gene, 193
 - hypothesis testing, 182–183, 192–195, 199
- Degrees of freedom, 28

- Dendrogram, 146
 - drawbacks, 163
 - labelled, 156
- Deoxyribonucleic acid, *see* DNA
- Design, 23
 - efficiency, 46
 - principles, 36–40
 - single-channel arrays, 40, 44
 - two channel arrays, 44–56
 - ratios, 125–129
 - variation, 24
- Design matrix, 45
- Diagnostic tests
 - designs for, 55
- Differential expression, 177
- Differential expression score, 178, 197
- Differentially expressed, 177
- Dimension reduction, 111–112
- Direct comparisons, 128
- Dissimilarity measures, 106–110
 - and measurement scale, 107
 - between clusters, 141–142
 - comparisons of, 108–110
 - correlation, 107, 161
 - Euclidean, 107, 140
 - for co-expression, 160–163
 - for repeated observations, 142–144
 - geometric, 140
 - Manhattan, 107, 140
 - missing data, 66–68
 - power distances, 107
 - weighted, 140–141, 143
- DNA (deoxyribonucleic acid), 15
- Dye balance, 40
- Dye effect normalization, 80–84
 - confounding with spatial correction, 132–133
 - constant, 80
 - dye-swap, 81
 - intensity dependent, 82–84
- Dye-swap
 - myths, 129
 - versus dye balance, 40
- Dye-swap normalization, *see* dye effect normalization, dye-swap
- dye.norm, 84
- Dyes, *see* Cy3/Cy5 dyes, cDNA labelling
- EMMIX-GENE, 174
- Enzymes, 18–19
- Error rates in hypothesis testing, 181–182
- Euclidean distance, *see* dissimilarity measures
- Expected misclassification rate, 213
- Expression profile, 160
- False array plots, 112
- False discovery rate (FDR), 182, 193
 - controlling the FDR, 195, 199
 - myths, 248–249
- False negative rate (FNR), 181
- False positive rate (FPR), 178, 181, 182
 - myths, 248
- Familywise error rate (FWER), 181, 193
 - controlling FWER, 193–195
 - myths, 247
- FDR, *see* false discovery rate
- Features, 104
- Filtering, 58–60
- Fixed circle segmentation, 61
- Flags, reliability, 66, 103
- FNR, *see* false negative rate
- FPR, *see* false positive rate
- FWER, *see* familywise error rate
- Gaussian distribution (normal distribution), 129–131

- Gene filtering, 218
 - biological selection, 222–223
 - clustering, 220–221
 - expression significance, 218–220
 - PCA, 220
 - penalized methods, 222
- Gene shaving, 170–171
- Generalized log transformations of scale, 64
- GEPAS, 68
- Gibbs sampling, 205
- Global error likelihood ratio test, 191–192
- Gridding, 60–61
- Guanine, 15

- hclust, 147
- Hierarchical clustering, 145–148
- Hierarchical PAM clustering, 150–155
 - global validation, 152–154
 - local validation, 153–154
- HIPAM, *see* hierarchical PAM clustering
- hipam, 154
- Histogram segmentation, 62
- Hochberg's FWER procedure, 194
- Hotelling's T^2 test, 219
- Housekeeping genes, 65
- Hybridization, 16, 20–21
 - buffer, 20
 - quality assessment of, 119–121, 124
- Hypothesis testing, 178
 - Bayesian, 196–210
 - classical, 179–196
 - comparison of methods, 195–196
 - decision rules, 182–183
- Image analysis, 57–62
- Imputation of missing data, 68–69
- impute.missing, 68
- Inactive genes, 177
- Indirect comparisons, 128
- Interwoven loop designs, 40, 49, 51–55
- Invariant set
 - across condition normalization, 89–92, 98
 - dye effect normalization, 82–84
- k -means clustering, 151
- k -nearest neighbours, *see* K-NN
- K-NN (k -nearest neighbours)
 - classification, 231–234
 - missing data imputation, 68–69
 - regression, 243–245
 - stepwise, 233, 243–244
- Kullback-Leibler divergence, 143–144

- Labelled PAM, 157–160
- Landing lights, 65
- lars, 241–243
- LASSO regression, 222, 237–243
- Layout of genes on array, 113, 120
- LDA, *see* linear discriminant analysis
- Learning set, *see* training set
- Least squares estimation, 235
- Likelihood, 191, 204
- Likelihood ratio test, 191–192
- Linear discriminant analysis
 - stepwise, 228–229
- Linear discriminant analysis (LDA), 227–231
- Linearity of expression data, 63
- Linkage, 141
 - average, 142
 - centroid, 142
 - complete, 142
 - median, 142
 - single, 142
- Locally weighted regression, *see* loess
- Loess, 75–76
 - smoothing, 73, 82

- Log ratios, 125–129
 - normalization, 69, 72
- Logarithmic scale for data, 64
- Mammary gland experiment, 9–10
- Manhattan distance, *see*
 - dissimilarity measures
- Mann-Whitney statistic, *see*
 - Wilcoxon rank sum statistic
- McLust, 172
- MDS, *see* multi-dimensional scaling
- Mean average silhouette width, 154
- Median filter, 58
- Medoid, 168
- Meta-genes, 220
- MGED, 104
- MIAME, 104–105
- Microarray, 15
- Minimum information about a
 - microarray experiment, *see* MIAME
- Misclassification costs, 212
- Misclassification rate, 212–215
 - cross-validation, 217
- Mishandling problems, 121–122, 124
- Mismatch probe, Affymetrix, 93
- Mismatch subtraction normalization, 94, 97
- Missing data, 65–69
- Misspecification rate, 225
- Mixture model
 - clustering, 171–175
 - hypothesis testing, 208–210
- MM, *see* mismatch probe
- Model-based clustering, 171–175
- Model-based expression index (MBEI), 100
- Morphological opening, 58
- Mouse mammary gland experiment, *see* mammary gland experiment
- Multidimensional scaling (MDS), 148
- Multiple testing decision rules, 193–195, 199
- Multiple testing problem, 179
- Nested design, 42
- Normal distribution (Gaussian distribution), 129–131
- Normalization, 62
 - across arrays, 84–93, 98–99
 - combining all effects, 39
 - one-channel arrays, 93–101
 - order of, 70–71
 - quality assessment of, 117–119, 123
 - sequential, 63, 69–101
 - two-channel arrays, 69–93
- Normalization and design, 39–40
- Nucleotides, 16
- Null hypothesis, 178, 180
- od, 50, 55
- Oligonucleotide, 93
- op, 33
- Optimal design, 44
 - A-optimality, 44, 47–55
 - weighted, 54–55
 - D-optimality, 44, 47–55
 - principles, 46–48
- Optimal pooling, *see* pooling samples, optimal pooling
- Outliers, 66
- Over-expressed, 178
- p -value, 178, 182–186, 188–190, 192, 193
- Pairwise scatterplots, 113
- PAM, *see* partitioning around medoids
- PAMSAM, 167–169
- pamsam, 169
- Partitioning around medoids (PAM), 151
 - labelled, 157–160
 - two-way, 165–167

- PCA, *see* principal component analysis
- PCA clustering, 148–150, *see* SVD clustering
- Pearson correlation, 161
- Penalized regression, 235–243
- Perfect match probe, Affymetrix, 93
- Permutation test (randomization test), 186–188
- Plaid models, 165
- `plot.lars`, 241
- `plot.tree`, 155
- PM, *see* perfect match probe
- Poly-A tail, 18
- Poly-T primer, 18
- Poly-T tail, 18
- Pooled variance *t*-statistic, *see t*-statistic
- Pooling samples, 30–36
 - arguments against, 30
 - biological variation, 31, 33–34
 - optimal pooling, 31–36
 - replication, 30
 - technical variation, 31, 34
- Posterior distribution, 205
- Posterior probability, 178
- Power of a test, 193
- Power transformations, 106
- Prediction
 - class membership, 223–234
 - continuous responses, 235–245
 - variance vs bias, 223–226, 235
- Predictor evaluation, 215
- Principal component analysis (PCA), 111, 112
- Print-tip spatial normalization, 72
- Prior distributions, 204
- Probabilistic background correction, 64, 78–80, 97
- Probe set, 99
 - summarization, 99–101
- Probes, 15
 - Affymetrix, 93
- q*-value, 199
- Quality assessment, 103
 - consequences, 123–124
 - qualitative, 104
- Quantification, 62
- Quantile normalization, 92, 99
 - across-condition, 89–92, 98
 - within-replication, 86–88, 98
- R computing environment, 1–3
- Randomization in design, 37–39
- Randomization test, *see* permutation test
- Reference designs, 51, 56, 72, 125
- Replication, 26, 38
 - biological replicates, 27–29
 - number of replicates, 29
 - on the same array, 28, 42
 - technical replicates, 27–29, 42–43
- Reporter*, 104
- Resistor-average, 143
- Reverse transcriptase, 19
- Ribonucleic acid, *see* RNA
- Ridge regression, 237
- RNA (ribonucleic acid), 16
 - extraction, 17–18
 - spiking, 18
- Sammon mapping, 112, 149–150
- `sammon.plot`, 158
- Scale for data, 63–64
 - in quality assessment, 105–106
- Scaling factors (Affymetrix), *see* array effect normalization
- Segmentation, 61–62
- Significance of a test, 182
- Silhouette width, 152
- Singular value decomposition, *see* SVD
- Skin cancer experiment, 6–7
- `smida`, 2–3
- Smoothing spatial trend
 - location, 73
 - scale, 73
- Smoothing splines, 82

- spat.norm, 74, 75, 95–96
- Spatial correction normalization
 - confounding with dye effect, 132–133
 - single channel, 95–96
 - two-channel, 71–76
- Spearman’s rank correlation, 161
- Spiking controls, 65
- step.knn, 233, 244
- step.lda, 229
- Stray signal, 96
- ‘Stress’, 112
- Strong control of FWER, 195
- Student *t*-distribution, 183, 184
- Super-genes, 148, 170, 220
- Supervised learning
 - exploratory, 155–160
- SVD, 148
 - clustering, 169–171
- SVD clustering, *see* PCA clustering
- SVD imputation, 68
- SVD impute, 68
- Symmetric percentile *t* method, 185

- t*-statistic
 - Bayesian hypothesis testing, 197, 198, 203–210
 - pooled variance, 188–189
 - Welch, 183–188
- Test statistic, 178, 181
 - for two conditions, 183–192
- Thymine, 15
- Time course experiments
 - clustering, 162–163, 174–175
 - Time series dissimilarity measures, 162–163
- Top-hat filter, 58–60
- Training set, 214
- Treatment, 37
- Tuberculosis experiment, 10–11
- Tukey biweight estimate, 94, 100
- Two-way PAM, 165–167
- twoway.pam, 166

- Unaffected genes, 177
- Under-expressed, 178
- Unit, statistical, 37
- Units of comparison, 110
- Unreliable data, 66

- Validation set, 215
- Variance measures of reliability, 66
- Variance-bias trade off
 - log ratios, 127–129
 - prediction, 223–226, 235
- Visualization methods, 111–112
- vsn, 79

- Weak control of FWER, 195
- Weighted designs, *see* optimal design
- Welch *t*-statistic, *see* *t*-statistic
- Wilcoxon rank sum statistic
 - Bayesian hypothesis testing, 197–198, 202–203
 - hypothesis testing, 189–190
- Within-cluster sums of squares, 167