

## Subject Index

- $N$ -mode tensor, 80  
 $\ell_1$ -norm, 94, 291  
 $\ell_1$ -regularizer, 88  
 $n$ -mode tensor product, 80
- activation function, 47  
 admissibility, 518  
 Akaike's information criterion (AIC), 365  
 all temperatures method, 382  
 approximate global variational Bayesian solver (AGVBS), 269  
 ARD Tucker, 332  
 asymptotic normality, 342, 352  
 asymptotic notation, 344  
 automatic relevance determination (ARD), 10, 36, 72, 204, 243, 294  
 average generalization error, 347  
 average training error, 347
- Bachmann–Landau notation, 344  
 backfitting algorithm, 283  
 basis selection effect, 373, 426  
 Bayes free energy, 36, 194, 348  
 Bayes generalization loss, 379  
 Bayes posterior, 4  
 Bayes theorem, 4  
 Bayes training loss, 379  
 Bayesian estimator, 5  
 Bayesian information criterion (BIC), 366  
 Bayesian learning, 3, 5  
 Bayesian network, 115, 455  
 Bernoulli distribution, 11, 27, 248, 253  
 Bernoulli mixture model, 451  
 Beta distribution, 11, 27  
 Beta function, 27  
 binomial distribution, 11, 27
- black-box variational inference, 50  
 burn-in, 59
- calculus of variations, 44  
 centering, 73  
 central limit theorem, 344  
 chi-squared distribution, 356  
 classification, 47  
 collaborative filtering (CF), 74, 335  
 collapsed Gibbs sampling, 53  
 collapsed MAP learning, 53  
 collapsed variational Bayesian learning, 53  
 complete likelihood, 8  
 conditional conjugacy, 39, 42  
 conditional distribution, 3  
 conditionally conjugate prior, 42  
 conjugacy, 10, 12  
 consistency, 352  
 continuation method, 267  
 convergence in distribution, 344  
 convergence in law, 344  
 convergence in probability, 344  
 coordinate descent, 66  
 core tensor, 80  
 Cramér–Rao lower-bound, 348  
 credible interval, 32  
 cross-validation, 9  
 cross-covariance, 73
- density, 522  
 Dirac delta function, 37, 52  
 direct site bounding, 49, 132  
 Dirichlet distribution, 11, 26  
 Dirichlet process prior, 112  
 distinct signal assumption, 419

- domination, 38, 518  
 doubly stochastic, 154
- efficiency, 518  
 empirical Bayesian (EBayes) estimator, 38, 193, 516  
 empirical Bayesian (EBayes) learning, 9, 35  
 empirical entropy, 349  
 empirical MAP (EMAP) learning, 311  
 empirical PB (EPB) learning, 311  
 empirical variational Bayesian (EVB) learning, 47  
 entropy, 380, 520  
 equivalence class, 187  
 error function, 47  
 Euler–Lagrange equation, 44  
 evidence, 36  
 evidence lower-bound (ELBO), 40, 341  
 exact global variational Bayesian solver (EGVBS), 267  
 expectation propagation (EP), 53  
 expectation-maximization (EM) algorithm, 9, 53, 105  
 exponential family, 15, 108, 443
- factor matrix, 80  
 finite mixture model, 103  
 Fisher information, 51, 197, 342, 520  
 foreground/background video separation, 97  
 forward–backward algorithm, 122  
 free energy, 40  
 free energy coefficient, 349
- Gamma distribution, 11, 16  
 Gamma function, 525  
 Gauss–Wishart distribution, 21  
 Gaussian distribution, 11  
 Gaussian mixture model, 104, 434  
 generalization coefficient, 348  
 generalization error, 335, 347  
 generalized Bayesian learning, 378  
 generalized posterior distribution, 378  
 generalized predictive distribution, 379  
 Gibbs generalization loss, 379  
 Gibbs learning, 380  
 Gibbs sampling, 59  
 Gibbs training loss, 379  
 global latent variable, 7, 103
- Hadamard product, 154  
 hard assignment, 9  
 hidden Markov model, 119, 461  
 hidden variable, 6  
 hierarchical model, 17  
 histogram, 26  
 homotopy method, 267  
 hyperparameter, 9  
 hyperprior, 9
- identifiability, 342, 352  
 improper prior, 200  
 independent and identically distributed (i.i.d.), 4  
 information criterion, 364  
 inside–outside algorithm, 126  
 integration effect, 374, 427  
 inverse temperature parameter, 379  
 isotropic Gauss–Gamma distribution, 17  
 isotropic Gaussian distribution, 11  
 iterative singular value shrinkage, 248, 252
- James–Stein (JS) estimator, 38, 516  
 Jeffreys prior, 198, 522  
 joint distribution, 3
- Kronecker delta, 130  
 Kronecker product, 66, 81, 331  
 Kronecker product covariance approximation (KPCA), 93, 274  
 Kullback–Leibler (KL) divergence, 39, 197, 347, 520
- Laplace approximation (LA), 51, 230  
 large-scale limit, 215, 319  
 latent Dirichlet allocation, 26, 127, 470  
 latent variable, 6  
 latent variable model, 103, 429  
 law of large numbers, 344  
 likelihood ratio, 347  
 linear neural network, 385  
 linear regression model, 22  
 link function, 253  
 local latent variable, 7, 103  
 local variational approximation, 49, 132  
 local-EMAP estimator, 317  
 local-EPB estimator, 317  
 local-EVB estimator, 182, 231, 242  
 log marginal likelihood, 36  
 log-concave distribution, 218  
 logistic regression, 132

- low-rank representation, 88  
 low-rank subspace clustering (LRSC), 88, 255
- Marčenko–Pastur (MP) distribution, 215  
 Marčenko–Pastur upper limit (MPUL), 216, 319
- marginal likelihood, 5  
 Markov chain Monte Carlo (MCMC), 58  
 matrix factorization (MF), 63, 195  
 matrix variate Gaussian, 66  
 maximum a posteriori (MAP) estimator, 188  
 maximum a posteriori (MAP) learning, 5, 294  
 maximum likelihood (ML) estimator, 188, 432  
 maximum likelihood (ML) learning, 5, 105  
 maximum log-likelihood, 366  
 mean update (MU) algorithm, 283  
 mean value theorem, 353  
 metric, 522  
 Metropolis–Hastings sampling, 58  
 minimum description length (MDL), 366  
 mixing weight, 7  
 mixture coefficient, 7  
 mixture model, 26, 196  
 mixture of Gaussians, 104  
 model distribution, 3  
 model likelihood, 3  
 model parameter, 3  
 model selection, 364  
 model-induced regularization (MIR), 72, 89, 94, 184, 195, 285, 308, 344, 373, 427  
 moment matching, 54  
 multilayer neural network, 196  
 multinomial distribution, 11, 26  
 multinomial parameter, 26
- natural parameter, 15, 108  
 neural network, 47  
 Newton–Raphson method, 108  
 noise variance parameter, 22  
 noninformative prior, 522  
 nonsingular, 23  
 normalization constant, 10  
 normalized cuts, 88
- Occam’s razor, 201, 366  
 one-of- $K$  representation, 8, 104, 455  
 overfitting, 420  
 overlap (OL) method, 230, 242
- Parafac, 80  
 partially Bayesian (PB) learning, 51, 131, 294
- partitioned-and-rearranged (PR) matrix, 95, 279  
 plug-in predictive distribution, 6, 46  
 Poisson distribution, 248, 253  
 polygamma function, 108  
 polynomial system, 162, 257, 267  
 positive-part James–Stein (PJS) estimator, 185, 307, 390, 518  
 posterior covariance, 5  
 posterior distribution, 4  
 posterior mean, 5  
 predictive distribution, 6  
 prior distribution, 3  
 probabilistic context-free grammar, 123, 466  
 probabilistic latent semantic analysis (pLSA), 131  
 probabilistic principal component analysis (probabilistic PCA), 63, 71, 230
- quasiconvexity, 207, 305
- radial basis function (RBF), 367  
 random matrix theory, 214, 375, 404  
 realizability, 346, 352  
 realizable, 375  
 rectified linear unit (ReLU), 47  
 reduced rank regression (RRR), 72, 385  
 regression parameter, 22  
 regular learning theory, 351  
 regular model, 198, 342  
 regularity condition, 342, 351  
 relative Bayes free energy, 348  
 relative variational Bayesian (VB) free energy, 383  
 resolution of singularities, 378  
 robust principal component analysis (robust PCA), 93, 288
- sample mean, 13  
 score function, 50  
 segmentation-based SAMF (sSAMF), 289  
 selecting the optimal basis function, 372  
 self-averaging, 216  
 sigmoid function, 47, 248, 253  
 simple variational Bayesian (SimpleVB) learning, 71  
 singular learning theory (SLT), 376  
 singular model, 197, 342, 522  
 singularities, 197, 342  
 soft assignment, 9

- sparse additive matrix factorization (SAMF),  
   94, 96, 204, 279  
 sparse matrix factorization (SMF) term, 94,  
   204, 279  
 sparse subspace clustering, 88  
 sparsity-inducing prior, 135  
 spectral clustering algorithm, 88  
 spiked covariance (SC) distribution, 217  
 spiked covariance model, 214  
 standard  $(K - 1)$ -simplex, 7  
 state density, 376  
 stick-breaking process, 112  
 stochastic complexity, 36  
 stochastic gradient descent, 50  
 strictly quasiconvex, 207  
 strong unimodality, 218  
 subspace clustering, 87  
 subtle signal assumption, 420  
 sufficient statistics, 15  
 superdiagonal, 80  
  
 Taylor approximation, 51  
 tensor, 80  
 tensor mode, 80  
 tensor rank, 80  
 trace norm, 88, 94, 291  
 training coefficient, 348  
  
 training error, 347  
 trial distribution, 39  
 Tucker factorization (TF), 80, 294, 331, 336  
  
 underfitting, 420  
 unidentifiability, 184, 187, 195  
 uniform prior, 198, 522  
 unnormalized posterior distribution, 4  
  
 variation, 44  
 variational Bayesian (VB) estimator, 46  
 variational Bayesian (VB) free energy, 383  
 variational Bayesian (VB) learning, 39  
 variational Bayesian (VB) posterior, 43, 46  
 variational parameter, 46, 66  
 vectorization operator, 66, 81, 331  
 volume element, 197, 522  
  
 weak convergence, 344  
 whitening, 73, 386  
 widely applicable Bayesian information  
   criterion (WBIC), 60, 381  
 widely applicable information criterion  
   (WAIC), 380  
 Wishart distribution, 11, 20  
  
 zeta function, 376