

- Byran, K. and T. Leise (2006). The 25,000,000,000 Eigenvector: The Linear Algebra behind Google. *SIAM Review* 48(3).
- Calvetti, D. and E. Somersalo (2007). *Introduction to Bayesian Scientific Computing*. Springer.
- Candes, E., J. Romberg, and T. Tao (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory* 52(2), 489–509.
- Candes, E. and M. Wakin (2008, March). An introduction to compressive sampling. *IEEE Signal Processing Magazine* 21.
- Candes, E., M. Wakin, and S. Boyd (2008). Enhancing sparsity by reweighted ℓ_1 minimization. *J. of Fourier Analysis and Applications* 1, 877–905.
- Cannings, C., E. A. Thompson, and M. H. Skolnick (1978). Probability functions in complex pedigrees. *Advances in Applied Probability* 10, 26–61.
- Canny, J. (2004). Gap: a factor model for discrete data. In *Proc. Annual Intl. ACM SIGIR Conference*, pp. 122–129.
- Cao, Z., T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li (2007). Learning to rank: From pairwise approach to listwise approach. In *Intl. Conf. on Machine Learning*, pp. 129–136.
- Cappe, O. (2010). Online Expectation Maximisation. In K. Mengersen, M. Titterton, and C. Robert (Eds.), *Mixtures*.
- Cappe, O. and E. Mouline (2009, June). Online EM Algorithm for Latent Data Models. *J. of Royal Stat. Soc. Series B* 71(3), 593–613.
- Cappe, O., E. Moulines, and T. Ryden (2005). *Inference in Hidden Markov Models*. Springer.
- Carbonetto, P. (2003). Unsupervised statistical models for general object recognition. Master's thesis, University of British Columbia.
- Carlin, B. P. and T. A. Louis (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall.
- Caron, F. and A. Doucet (2008). Sparse Bayesian nonparametric regression. In *Intl. Conf. on Machine Learning*.
- Carreira-Perpinan, M. and C. Williams (2003). An isotropic gaussian mixture can have more modes than components. Technical Report EDI-INF-RR-0185, School of Informatics, U. Edinburgh.
- Carter, C. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika* 81(3), 541–553.
- Carterette, B., P. Bennett, D. Chickering, and S. Dumais (2008). Here or There: Preference Judgments for Relevance. In *Proc. ECIR*.
- Caruana, R. (1998). A dozen tricks with multitask learning. In G. Orr and K.-R. Mueller (Eds.), *Neural Networks: Tricks of the Trade*. Springer-Verlag.
- Caruana, R. and A. Niculescu-Mizil (2006). An empirical comparison of supervised learning algorithms. In *Intl. Conf. on Machine Learning*.
- Carvahlo, C., N. Polson, and J. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika* 97(2), 465.
- Carvahlo, L. and C. Lawrence (2007). Centroid estimation in discrete high-dimensional spaces with applications in biology. *Proc. of the National Academy of Science, USA* 105(4).
- Carvalho, C. M. and M. West (2007). Dynamic matrix-variate graphical models. *Bayesian Analysis* 2(1), 69–98.
- Casella, G. and R. Berger (2002). *Statistical inference*. Duxbury. 2nd edition.
- Castro, M., M. Coates, and R. D. Nowak (2004). Likelihood based hierarchical clustering. *IEEE Trans. in Signal Processing* 52(8), 230.
- Celeux, G. and J. Diebolt (1985). The SEM algorithm: A probabilistic teacher derive from the EM algorithm for the mixture problem. *Computational Statistics Quarterly* 2, 73–82.
- Cemgil, A. T. (2001). A technique for painless derivation of kalman filtering recursions. Technical report, U. Nijmegen.
- Cesa-Bianchi, N. and G. Lugosi (2006). *Prediction, learning, and games*. Cambridge University Press.
- Cevher, V. (2009). Learning with compressible priors. In *NIPS*.
- Chai, K. M. A. (2010). *Multi-task learning with Gaussian processes*. Ph.D. thesis, U. Edinburgh.
- Chang, H., Y. Weiss, and W. Freeman (2009). Informative Sensing. Technical report, Hebrew U. Submitted to IEEE Transactions on Info. Theory.
- Chang, J. and D. Blei (2010). Hierarchical relational models for document networks. *The Annals of Applied Statistics* 4(1), 124–150.
- Chang, J., J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei (2009). Reading tea leaves: How humans interpret topic models. In *NIPS*.
- Chapelle, O. and L. Li (2011). An empirical evaluation of Thompson sampling. In *NIPS*.
- Chartrand, R. and W. Yin (2008). Iteratively reweighted algorithms for compressive sensing. In *Intl. Conf. on Acoustics, Speech and Signal Proc.*
- Chechik, G., A. G. N. Tishby, and Y. Weiss (2005). Information bottleneck for gaussian variables. *J. of Machine Learning Research* 6, 165–188.
- Cheeseman, P., J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman (1988). Autoclass: A Bayesian classification system. In *Proc. of the Fifth Intl. Workshop on Machine Learning*.
- Cheeseman, P. and J. Stutz (1996). Bayesian classification (autoclass): Theory and results. In Fayyad, Prateksky-Shapiro, Smyth, and Uthuramy (Eds.), *Advances in Knowledge Discovery and Data Mining*. MIT Press.
- Chen, B., K. Swersky, B. Marlin, and N. de Freitas (2010). Sparsity priors and boosting for learning localized distributed feature representations. Technical report, UBC.
- Chen, B., J.-A. Ting, B. Marlin, and N. de Freitas (2010). Deep learning of invariant spatio-temporal features from video. In *NIPS Workshop on Deep Learning*.

- Chen, M., D. Carlson, A. Zaas, C. Woods, G. Ginsburg, A. Hero, J. Lucas, and L. Carin (2011, March). The Bayesian Elastic Net: Classifying Multi-Task Gene-Expression Data. *IEEE Trans. Biomed. Eng.* 58(3), 468–79.
- Chen, R. and S. Liu (2000). Mixture Kalman filters. *J. Royal Stat. Soc. B.*
- Chen, S. and J. Goodman (1996). An empirical study of smoothing techniques for language modeling. In *Proc. 34th ACL*, pp. 310–318.
- Chen, S. and J. Goodman (1998). An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Dept. Comp. Sci., Harvard.
- Chen, S. and J. Wigger (1995, July). Fast orthogonal least squares algorithm for efficient subset model selection. *IEEE Trans. Signal Processing* 3(7), 1713–1715.
- Chen, S. S., D. L. Donoho, and M. A. Saunders (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* 20(1), 33–61.
- Chen, X., S. Kim, Q. Lin, J. G. Carbonell, and E. P. Xing (2010). Graph-Structured Multi-task Regression and an Efficient Optimization Method for General Fused Lasso. Technical report, CMU.
- Chib, S. (1995). Marginal likelihood from the Gibbs output. *J. of the Am. Stat. Assoc.* 90, 1313–1321.
- Chickering, D. (1996). Learning Bayesian networks is NP-Complete. In *AI/Stats V.*
- Chickering, D. and D. Heckerman (1997). Efficient approximations for the marginal likelihood of incomplete data given a Bayesian network. *Machine Learning* 29, 181–212.
- Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3, 507–554.
- Chipman, H., E. George, and R. McCulloch (1998). Bayesian CART model search. *J. of the Am. Stat. Assoc.* 93, 935–960.
- Chipman, H., E. George, and R. McCulloch (2001). The practical implementation of Bayesian Model Selection. *Model Selection*. IMS Lecture Notes.
- Chipman, H., E. George, and R. McCulloch (2006). Bayesian Ensemble Learning. In *NIPS*.
- Chipman, H., E. George, and R. McCulloch (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* 4(1), 266–298.
- Choi, M., V. Tan, A. Anandkumar, and A. Willsky (2011). Learning latent tree graphical models. *J. of Machine Learning Research*.
- Choi, M. J. (2011). *Trees and Beyond: Exploiting and Improving Tree-Structured Graphical Models*. Ph.D. thesis, MIT.
- Choset, H. and K. Nagatani (2001). Topological simultaneous localization and mapping (SLAM): toward exact localization without explicit localization. *IEEE Trans. Robotics and Automation* 17(2).
- Chow, C. K. and C. N. Liu (1968). Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory* 14, 462–67.
- Christensen, O., G. Roberts, and M. SkÅld (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *J. of Computational and Graphical Statistics* 15, 1–17.
- Chung, F. (1997). *Spectral Graph Theory*. AMS.
- Cimiano, P., A. Schultz, S. Sizov, P. Sorg, and S. Staab (2009). Explicit versus latent concept models for cross-language information retrieval. In *Intl. Joint Conf. on AI*.
- Cipra, B. (2000). The Ising Model Is NP-Complete. *SIAM News* 33(6).
- Ciresan, D. C., U. Meier, L. M. Gambardella, and J. Schmidhuber (2010). Deep big simple neural nets for handwritten digit recognition. *Neural Computation* 22(12), 3207–3220.
- Clarke, B. (2003). Bayes model averaging and stacking when model approximation error cannot be ignored. *J. of Machine Learning Research*, 683–712.
- Clarke, B., E. Fokoue, and H. H. Zhang (2009). *Principles and Theory for Data Mining and Machine Learning*. Springer.
- Cleveland, W. and S. Devlin (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *J. of the Am. Stat. Assoc.* 83(403), 596–610.
- Collins, M. (2002). Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *EMNLP*.
- Collins, M., S. Dasgupta, and R. E. Schapire (2002). A generalization of principal components analysis to the exponential family. In *NIPS-14*.
- Collins, M. and N. Duffy (2002). Convolution kernels for natural language. In *NIPS*.
- Collobert, R. and J. Weston (2008). A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. In *Intl. Conf. on Machine Learning*.
- Combettes, P. and V. Wajs (2005). Signal recovery by proximal forward-backward splitting. *SIAM J. Multi-scale Model. Simul.* 4(4), 1168–1200.
- Cook, J. (2005). Exact Calculation of Beta Inequalities. Technical report, M. D. Anderson Cancer Center, Dept. Biostatistics.
- Cooper, G. and E. Herskovits (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9, 309–347.
- Cooper, G. and C. Yoo (1999). Causal discovery from a mixture of experimental and observational data. In *UAI*.
- Cover, T. and P. Hart (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* 13(1), 21–27.
- Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory*. John Wiley.
- Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory*. John Wiley. 2nd edition.
- Cowles, M. and B. Carlin (1996). Markov chain monte carlo convergence diagnostics: A comparative review. *J. of the Am. Stat. Assoc.* 91, 883–904.

- Crisan, D., P. D. Moral, and T. Lyons (1999). Discrete filtering using branching and interacting particle systems. *Markov Processes and Related Fields* 5(3), 293–318.
- Cui, Y., X. Z. Fern, and J. G. Dy (2010). Learning multiple nonredundant clusterings. *ACM Transactions on Knowledge Discovery from Data* 4(3).
- Cukier, K. (2010, February). Data, data everywhere.
- Dagum, P. and M. Luby (1993). Approximating probabilistic inference in Bayesian belief networks is NP-hard. *Artificial Intelligence* 60, 141–153.
- Dahl, J., L. Vandenberghe, and V. Roychowdhury (2008, August). Covariance selection for non-chordal graphs via chordal embedding. *Optimization Methods and Software* 23(4), 501–502.
- Dahlhaus, R. and M. Eichler (2000). Causality and graphical models for time series. In P. Green, N. Hjort, and S. Richardson (Eds.), *Highly structured stochastic systems*. Oxford University Press.
- Dallal, S. and W. Hall (1983). Approximating priors by mixtures of natural conjugate priors. *J. of Royal Stat. Soc. Series B* 45, 278–286.
- Darwiche, A. (2009). *Modeling and Reasoning with Bayesian Networks*. Cambridge.
- Daume, H. (2007a). Fast search for Dirichlet process mixture models. In *AI/Statistics*.
- Daume, H. (2007b). Frustratingly easy domain adaptation. In *Proc. the Assoc. for Comp. Ling.*
- Dawid, A. P. (1992). Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing* 2, 25–36.
- Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *Intl. Stat. Review* 70, 161–189. Corrections p437.
- Dawid, A. P. (2010). Beware of the DAG! *J. of Machine Learning Research* 6, 59–86.
- Dawid, A. P. and S. L. Lauritzen (1993). Hyper-markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics* 3, 1272–1317.
- de Freitas, N., R. Dearden, F. Hutter, R. Morales-Menendez, J. Mutch, and D. Poole (2004). Diagnosis by a waiter and a mars explorer. *Proc. IEEE* 92(3).
- de Freitas, N., M. Niranjan, and A. Gee (2000). Hierarchical Bayesian models for regularisation in sequential learning. *Neural Computation* 12(4), 955–993.
- Dechter, R. (1996). Bucket elimination: a unifying framework for probabilistic inference. In *UAI*.
- Dechter, R. (2003). *Constraint Processing*. Morgan Kaufmann.
- Decoste, D. and B. Schoelkopf (2002). Training invariant support vector machines. *Machine learning* 41, 161–190.
- Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *J. of the American Society for Information Science* 41(6), 391–407.
- DeGroot, M. (1970). *Optimal Statistical Decisions*. McGraw-Hill.
- Deisenroth, M., C. Rasmussen, and J. Peters (2009). Gaussian Process Dynamic Programming. *Neurocomputing* 72(7), 1508–1524.
- Dellaportas, P., P. Giudici, and G. Roberts (2003). Bayesian inference for nondecomposable graphical gaussian models. *Sankhya, Ser. A* 65, 43–55.
- Dellaportas, P. and A. F. M. Smith (1993). Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling. *J. of the Royal Statistical Society. Series C (Applied Statistics)* 42(3), 443–459.
- Delyon, B., M. Lavielle, and E. Moulines (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics* 27(1), 94–128.
- Dempster, A. (1972). Covariance selection. *Biometrics* 28(1).
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Statistical Society, Series B* 34, 1–38.
- Denison, D., C. Holmes, B. Mallick, and A. Smith (2002). *Bayesian methods for nonlinear classification and regression*. Wiley.
- Denison, D., B. Mallick, and A. Smith (1998). A Bayesian CART algorithm. *Biometrika* 85, 363–377.
- Desjardins, G. and Y. Bengio (2008). Empirical evaluation of convolutional RBMs for vision. Technical Report 1327, U. Montreal.
- Dey, D., S. Ghosh, and B. Mallick (Eds.) (2000). *Generalized Linear Models: A Bayesian Perspective*. Chapman & Hall/CRC Biostatistics Series.
- Diaconis, P., S. Holmes, and R. Montgomery (2007). Dynamical Bias in the Coin Toss. *SIAM Review* 49(2), 211–235.
- Diaconis, P. and D. Ylvisaker (1985). Quantifying prior opinion. In *Bayesian Statistics 2*.
- Dietterich, T. G. and G. Bakiri (1995). Solving multiclass learning problems via ECOCs. *J. of AI Research* 2, 263–286.
- Diggle, P. and P. Ribeiro (2007). *Model-based Geostatistics*. Springer.
- Ding, Y. and R. Harrison (2010). A sparse multinomial probit model for classification. *Pattern Analysis and Applications*, 1–9.
- Dobra, A. (2009). Dependency networks for genome-wide data. Technical report, U. Washington.
- Dobra, A. and H. Massam (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Statistical Methodology* 7, 240–253.
- Domingos, P. and D. Lowd (2009). *Markov Logic: An Interface Layer for AI*. Morgan & Claypool.
- Domingos, P. and M. Pazzani (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning* 29, 103–130.

- Domke, J., A. Karapurkar, and Y. Aloimonos (2008). Who killed the directed model? In *CVPR*.
- Doucet, A., N. de Freitas, and N. J. Gordon (2001). *Sequential Monte Carlo Methods in Practice*. Springer Verlag.
- Doucet, A., N. Gordon, and V. Krishnamurthy (2001). Particle Filters for State Estimation of Jump Markov Linear Systems. *IEEE Trans. on Signal Processing* 49(3), 613–624.
- Dow, J. and J. Endersby (2004). Multinomial probit and multinomial logit: a comparison of choice models for voting research. *Electoral Studies* 23(1), 107–122.
- Drineas, P., A. Frieze, R. Kannan, S. Vempala, and V. Vinay (2004). Clustering large graphs via the singular value decomposition. *Machine Learning* 56, 9–33.
- Drugowitsch, J. (2008). Bayesian linear regression. Technical report, U. Rochester.
- Druilhet, P. and J.-M. Marin (2007). Invariant HPD credible sets and MAP estimators. *Bayesian Analysis* 2(4), 681–692.
- Duane, S., A. Kennedy, B. Pendleton, and D. Roweth (1987). Hybrid Monte Carlo. *Physics Letters B* 195(2), 216–222.
- Duchi, J., S. Gould, and D. Koller (2008). Projected subgradient methods for learning sparse gaussians. In *UAI*.
- Duchi, J., E. Hazan, and Y. Singer (2010). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. In *Proc. of the Workshop on Computational Learning Theory*.
- Duchi, J., S. Shalev-Shwartz, Y. Singer, and T. Chandra (2008). Efficient projections onto the L1-ball for learning in high dimensions. In *Intl. Conf. on Machine Learning*.
- Duchi, J. and Y. Singer (2009). Boosting with structural sparsity. In *Intl. Conf. on Machine Learning*.
- Duchi, J., D. Tarlow, G. Elidan, and D. Koller (2007). Using combinatorial optimization within max-product belief propagation. In *NIPS*.
- Duda, R. O., P. E. Hart, and D. G. Stork (2001). *Pattern Classification*. Wiley Interscience. 2nd edition.
- Dumais, S. and T. Landauer (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104, 211–240.
- Dunson, D., J. Palomo, and K. Bollen (2005). Bayesian Structural Equation Modeling. Technical Report 2005-5, SAMSI.
- Durbin, J. and S. J. Koopman (2001). *Time Series Analysis by State Space Methods*. Oxford University Press.
- Durbin, R., S. Eddy, A. Krogh, and G. Mitchison (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.
- Earl, D. and M. Deem (2005). Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* 7, 3910.
- Eaton, D. and K. Murphy (2007). Exact Bayesian structure learning from uncertain interventions. In *AI/Statistics*.
- Edakunni, N., S. Schaal, and S. Vijayakumar (2010). Probabilistic incremental locally weighted learning using randomly varying coefficient model. Technical report, USC.
- Edwards, D., G. de Abreu, and R. Labouriau (2010). Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. *BMC Bioinformatics* 11(18).
- Efron, B. (1986). Why Isn't Everyone a Bayesian? *The American Statistician* 40(1).
- Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge.
- Efron, B., I. Johnstone, T. Hastie, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics* 32(2), 407–499.
- Efron, B. and C. Morris (1975). Data analysis using stein's estimator and its generalizations. *J. of the Am. Stat. Assoc.* 70(350), 311–319.
- Elad, M. and I. Yavneh (2009). A plurality of sparse representations is better than the sparsest one alone. *IEEE Trans. on Info. Theory* 55(10), 4701–4714.
- Elidan, G. and S. Gould (2008). Learning Bounded Treewidth Bayesian Networks. *J. of Machine Learning Research*, 2699–2731.
- Elidan, G., N. Lotner, N. Friedman, and D. Koller (2000). Discovering hidden variables: A structure-based approach. In *NIPS*.
- Elidan, G., I. McGraw, and D. Koller (2006). Residual belief propagation: Informed scheduling for asynchronous message passing. In *UAI*.
- Elkan, C. (2003). Using the triangle inequality to accelerate k-means. In *Intl. Conf. on Machine Learning*.
- Elkan, C. (2005). Deriving TF-IDF as a Fisher kernel. In *Proc. Intl. Symp. on String Processing and Information Retrieval (SPIRE)*, pp. 296–301.
- Elkan, C. (2006). Clustering documents with an exponential family approximation of the Dirichlet compound multinomial model. In *Intl. Conf. on Machine Learning*.
- Ellis, B. and W. H. Wong (2008). Learning causal bayesian network structures from experimental data. *J. of the Am. Stat. Assoc.* 103(482), 778–789.
- Engel, Y., S. Mannor, and R. Meir (2005). Reinforcement Learning with Gaussian Processes. In *Intl. Conf. on Machine Learning*.
- Erhan, D., Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio (2010). Why Does Unsupervised Pre-training Help Deep Learning? *J. of Machine Learning Research* 11, 625–660.
- Erosheva, E., S. Fienberg, and C. Joutard (2007). Describing disability through individual-level mixture models for multivariate binary data. *Annals of Applied Statistics*.
- Erosheva, E., S. Fienberg, and J. Lafferty (2004). Mixed-membership models of scientific publications. *Proc. of the National Academy of Science, USA* 101, 5220–2227.

- Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *J. of the Am. Stat. Assoc.* 90(430), 577–588.
- Ewens, W. (1990). Population genetics theory - the past and the future. In S. Lessard (Ed.), *Mathematical and Statistical Developments of Evolutionary Theory*, pp. 177–227. Reidel.
- Fan, J. and R. Z. Li (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *J. of the Am. Stat. Assoc.* 96(456), 1348–1360.
- Fearnhead, P. (2004). Exact bayesian curve fitting and signal segmentation. *IEEE Trans. Signal Processing* 53, 2160–2166.
- Felzenszwalb, P. and D. Huttenlocher (2006). Efficient belief propagation for early vision. *Intl. J. Computer Vision* 70(1), 41–54.
- Ferrucci, D., E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. W. Murdock, E. N. amd J. Prager, N. Schlaefter, and C. Welty (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 59–79.
- Fienberg, S. (1970). An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics* 41(3), 907a–917.
- Figueiredo, M. (2003). Adaptive sparseness for supervised learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 25(9), 1150–1159.
- Figueiredo, M., R. Nowak, and S. Wright (2007). Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE J. on Selected Topics in Signal Processing*.
- Figueiredo, M. A. T. and A. K. Jain (2002). Unsupervised learning of finite mixture models. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24(3), 381–396. Matlab code at <http://www.lx.it.pt/~mtf/mixture-code.zip>.
- Fine, S., Y. Singer, and N. Tishby (1998). The hierarchical Hidden Markov Model: Analysis and applications. *Machine Learning* 32, 41.
- Finkel, J. and C. Manning (2009). Hierarchical bayesian domain adaptation. In *Proc. NAACL*, pp. 602–610.
- Fischer, B. and J. Schumann (2003). Autobayes: A system for generating data analysis programs from statistical models. *J. Functional Programming* 13(3), 483–508.
- Fishelson, M. and D. Geiger (2002). Exact genetic linkage computations for general pedigrees. *BMC Bioinformatics* 18.
- Fletcher, R. (2005). On the Barzilai-Borwein Method. *Applied Optimization* 96, 235–256.
- Fokoue, E. (2005). Mixtures of factor analyzers: an extension with covariates. *J. Multivariate Analysis* 95, 370–384.
- Forbes, J., T. Huang, K. Kanazawa, and S. Russell (1995). The BATmobile: Towards a Bayesian automated taxi. In *Intl. Joint Conf. on AI*.
- Forsyth, D. and J. Ponce (2002). *Computer vision: a modern approach*. Prentice Hall.
- Fraley, C. and A. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *J. of the Am. Stat. Assoc.* (97), 611–631.
- Fraley, C. and A. Raftery (2007). Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering. *J. of Classification* 24, 155–181.
- Franc, V., A. Zien, and B. Schoelkopf (2011). Support vector machines as probabilistic models. In *Intl. Conf. on Machine Learning*.
- Frank, I. and J. Friedman (1993). A statistical view of some chemometrics regression tools. *Technometrics* 35(2), 109–135.
- Fraser, A. (2008). *Hidden Markov Models and Dynamical Systems*. SIAM Press.
- Freund, Y. and R. R. Schapire (1996). Experiments with a new boosting algorithm. In *Intl. Conf. on Machine Learning*.
- Frey, B. (1998). *Graphical Models for Machine Learning and Digital Communication*. MIT Press.
- Frey, B. (2003). Extending factor graphs so as to unify directed and undirected graphical models. In *UAI*.
- Frey, B. and D. Dueck (2007, February). Clustering by Passing Messages Between Data Points. *Science* 315, 972a–976.
- Friedman, J. (1991). Multivariate adaptive regression splines. *Ann. Statist.* 19, 1–67.
- Friedman, J. (1997a). On bias, variance, 0-1 loss and the curse of dimensionality. *J. Data Mining and Knowledge Discovery* 1, 55–77.
- Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 29, 1189–1232.
- Friedman, J., T. Hastie, and R. Tibshirani (2000). Additive logistic regression: a statistical view of boosting. *Annals of statistics* 28(2), 337–374.
- Friedman, J., T. Hastie, and R. Tibshirani (2008). Sparse inverse covariance estimation the graphical lasso. *Biostatistics* 9(3), 432–441.
- Friedman, J., T. Hastie, and R. Tibshirani (2010, February). Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. of Statistical Software* 33(1).
- Friedman, N. (1997b). Learning Bayesian networks in the presence of missing values and hidden variables. In *UAI*.
- Friedman, N., D. Geiger, and M. Goldszmidt (1997). Bayesian network classifiers. *Machine Learning J.* 29, 131–163.
- Friedman, N., D. Geiger, and N. Lotner (2000). Likelihood computation with value abstraction. In *UAI*.
- Friedman, N. and D. Koller (2003). Being Bayesian about Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning* 50, 95–126.
- Friedman, N., M. Ninion, I. Pe'er, and T. Pupko (2002). A Structural EM Algorithm for Phylogenetic Inference. *J. Comp. Bio.* 9, 331–353.
- Friedman, N. and Y. Singer (1999). Efficient Bayesian parameter estimation in large discrete domains. In *NIPS-11*.

- Fruhwirth-Schnatter, S. (2007). *Finite Mixture and Markov Switching Models*. Springer.
- Fruhwirth-Schnatter, S. and R. Fruhwirth (2010). Data Augmentation and MCMC for Binary and Multinomial Logit Models. In T. Kneib and G. Tutz (Eds.), *Statistical Modelling and Regression Structures*, pp. 111–132. Springer.
- Fu, W. (1998). Penalized regressions: the bridge versus the lasso. *J. Computational and graphical statistics*.
- Fukushima, K. (1975). Cognitron: a self-organizing multilayered neural network. *Biological Cybernetics* 20(6), 121–136.
- Fung, R. and K. Chang (1989). Weighting and integrating evidence for stochastic simulation in Bayesian networks. In *UAI*.
- Gabow, H., Z. Galil, and T. Spencer (1984). Efficient implementation of graph algorithms using contraction. In *IEEE Symposium on the Foundations of Computer Science*.
- Gales, M. (2002). Maximum likelihood multiple subspace projections for hidden Markov models. *IEEE Trans. on Speech and Audio Processing* 10(2), 37–47.
- Gales, M. J. F. (1999). Semi-tied covariance matrices for hidden Markov models. *IEEE Trans. on Speech and Audio Processing* 7(3), 272–281.
- Gamerman, D. (1997). Efficient sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing* 7, 57–68.
- Geiger, D. and D. Heckerman (1994). Learning Gaussian networks. In *UAI*, Volume 10, pp. 235–243.
- Geiger, D. and D. Heckerman (1997). A characterization of Dirichlet distributions through local and global independence. *Annals of Statistics* 25, 1344–1368.
- Gelfand, A. (1996). Model determination using sampling-based methods. In Gilks, Richardson, and Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*. Chapman & Hall.
- Gelfand, A. and A. Smith (1990). Sampling-based approaches to calculating marginal densities. *J. of the Am. Stat. Assoc.* 85, 385–409.
- Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004). *Bayesian data analysis*. Chapman and Hall. 2nd edition.
- Gelman, A. and J. Hill (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge.
- Gelman, A. and X.-L. Meng (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science* 13, 163–185.
- Gelman, A. and T. Raghunathan (2001). Using conditional distributions for missing-data imputation. *Statistical Science*.
- Gelman, A. and D. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7, 457–511.
- Geman, S., E. Bienenstock, and R. Doursat (1992). Neural networks and the bias-variance dilemma. *Neural Computing* 4, 1–58.
- Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 6(6).
- Geoffrion, A. (1974). Lagrangian relaxation for integer programming. *Mathematical Programming Study* 2, 82–114.
- George, E. and D. Foster (2000). Calibration and empirical bayes variable selection. *Biometrika* 87(4), 731–747.
- Getoor, L. and B. Taskar (Eds.) (2007). *Introduction to Relational Statistical Learning*. MIT Press.
- Geyer, C. (1992). Practical markov chain monte carlo. *Statistical Science* 7, 473–483.
- Ghahramani, Z. and M. Beal (2000). Variational inference for Bayesian mixtures of factor analysers. In *NIPS-12*.
- Ghahramani, Z. and M. Beal (2001). Propagation algorithms for variational Bayesian learning. In *NIPS-13*.
- Ghahramani, Z. and G. Hinton (1996a). The EM algorithm for mixtures of factor analysers. Technical report, Dept. of Comp. Sci., Uni. Toronto.
- Ghahramani, Z. and G. Hinton (1996b). Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2, Dept. Comp. Sci., Univ. Toronto.
- Ghahramani, Z. and M. Jordan (1997). Factorial hidden Markov models. *Machine Learning* 29, 245–273.
- Gilks, W. and C. Berzuini (2001). Following a moving target – Monte Carlo inference for dynamic Bayesian models. *J. of Royal Stat. Soc. Series B* 63, 127–146.
- Gilks, W., N. Best, and K. Tan (1995). Adaptive rejection Metropolis sampling. *Applied Statistics* 44, 455–472.
- Gilks, W. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics* 41, 337–348.
- Girolami, M., B. Calderhead, and S. Chin (2010). Riemannian Manifold Hamiltonian Monte Carlo. *J. of Royal Stat. Soc. Series B*. To appear.
- Girolami, M. and S. Rogers (2005). Hierarchic bayesian models for kernel learning. In *Intl. Conf. on Machine Learning*, pp. 241–248.
- Girolami, M. and S. Rogers (2006). Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Computation* 18(8), 1790 – 1817.
- Girshick, R., P. Felzenszwalb, and D. McAllester (2011). Object detection with grammar models. In *NIPS*.
- Gittins, J. (1989). *Multi-armed Bandit Allocation Indices*. Wiley.
- Giudici, P. and P. Green (1999). Decomposable graphical gaussian model determination. *Biometrika* 86(4), 785–801.
- Givoni, I. E. and B. J. Frey (2009, June). A binary variable model for affinity propagation. *Neural Computation* 21(6), 1589–1600.
- Globerson, A. and T. Jaakkola (2008). Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *NIPS*.
- Glorot, X. and Y. Bengio (2010, May). Understanding the difficulty of training deep feedforward neural networks. In *AI/Statistics*, Volume 9, pp. 249–256.

- Gogate, V., W. A. Webb, and P. Domingos (2010). Learning efficient Markov networks. In *NIPS*.
- Goldenberg, A., A. X. Zheng, S. E. Fienberg, and E. M. Airoldi (2009). A Survey of Statistical Network Models. *Foundations and Trends in Machine Learning*, 129–233.
- Golub, G. and C. F. van Loan (1996). *Matrix computations*. Johns Hopkins University Press.
- Gonen, M., W. Johnson, Y. Lu, and P. Westfall (2005, August). The Bayesian Two-Sample t Test. *The American Statistician* 59(3), 252–257.
- Gonzales, T. (1985). Clustering to minimize the maximum intercluster distance. *Theor. Comp. Sci.* 38, 293–306.
- Gorder, P. F. (2006, Nov/Dec). Neural networks show new promise for machine vision. *Computing in science & engineering* 8(6), 4–8.
- Gordon, N. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEEE Proceedings (F)* 140(2), 107–113.
- Graepel, T., J. Quinero-Candela, T. Borchert, and R. Herbrich (2010). Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft’s Bing Search Engine. In *Intl. Conf. on Machine Learning*.
- Grauman, K. and T. Darrell (2007, April). The Pyramid Match Kernel: Efficient Learning with Sets of Features. *J. of Machine Learning Research* 8, 725–760.
- Green, P. (1998). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 82, 711–732.
- Green, P. (2003). Tutorial on trans-dimensional MCMC. In P. Green, N. Hjort, and S. Richardson (Eds.), *Highly Structured Stochastic Systems*. OUP.
- Green, P. and B. Silverman (1994). *Non-parametric regression and generalized linear models*. Chapman and Hall.
- Greenshtein, E. and J. Park (2009). Application of Non Parametric Empirical Bayes Estimation to High Dimensional Classification. *J. of Machine Learning Research* 10, 1687–1704.
- Greig, D., B. Porteous, and A. Seheult (1989). Exact maximum a posteriori estimation for binary images. *J. of Royal Stat. Soc. Series B* 51(2), 271–279.
- Griffin, J. and P. Brown (2007). Bayesian adaptive lassos with non-convex penalization. Technical report, U. Kent.
- Griffin, J. and P. Brown (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* 5(1), 171–188.
- Griffiths, T. and J. Tenenbaum (2009). Theory-Based Causal Induction. *Psychological Review* 116(4), 661–716.
- Griffiths, T. and M. Steyvers (2004). Finding scientific topics. *Proc. of the National Academy of Science, USA* 101, 5228–5235.
- Griffiths, T., M. Steyvers, D. Blei, and J. Tenenbaum (2004). Integrating topics and syntax. In *NIPS*.
- Griffiths, T. and J. Tenenbaum (2001). Using vocabulary knowledge in bayesian multinomial estimation. In *NIPS*, pp. 1385–1392.
- Griffiths, T. and J. Tenenbaum (2005). Structure and strength in causal induction. *Cognitive Psychology* 51, 334–384.
- Grimmett, G. and D. Stirzaker (1992). *Probability and Random Processes*. Oxford.
- Guan, Y., J. Dy, D. Niu, and Z. Ghahramani (2010). Variational Inference for Nonparametric Multiple Clustering. In *1st Intl. Workshop on Discovering, Summarizing and Using Multiple Clustering (MultiClust)*.
- Guedon, Y. (2003). Estimating hidden semi-markov chains from discrete sequences. *J. of Computational and Graphical Statistics* 12, 604–639.
- Guo, Y. (2009). Supervised exponential family principal component analysis via convex optimization. In *NIPS*.
- Gustafsson, M. (2001). A probabilistic derivation of the partial least-squares algorithm. *Journal of Chemical Information and Modeling* 41, 288–294.
- Guyon, I., S. Gunn, M. Nikravesh, and L. Zadeh (Eds.) (2006). *Feature Extraction: Foundations and Applications*. Springer.
- Hacker, J. and P. Pierson (2010). *Winner-Take-All Politics: How Washington Made the Rich Richer—and Turned Its Back on the Middle Class*. Simon & Schuster.
- Halevy, A., P. Norvig, and F. Pereira (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24(2), 8–12.
- Hall, P., J. T. Ormerod, and M. P. Wand (2011). Theory of Gaussian Variational Approximation for a Generalised Linear Mixed Model. *Statistica Sinica* 21, 269–389.
- Hamilton, J. (1990). Analysis of time series subject to changes in regime. *J. Econometrics* 45, 39–70.
- Hans, C. (2009). Bayesian Lasso regression. *Biometrika* 96(4), 835–845.
- Hansen, M. and B. Yu (2001). Model selection and the principle of minimum description length. *J. of the Am. Stat. Assoc.*
- Hara, H. and A. Takimura (2008). A Localization Approach to Improve Iterative Proportional Scaling in Gaussian Graphical Models. *Communications in Statistics - Theory and Method*. to appear.
- Hardin, J. and J. Hilbe (2003). *Generalized Estimating Equations*. Chapman and Hall/CRC.
- Harmeling, S. and C. K. I. Williams (2011). Greedy learning of binary latent trees. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 33(6), 1087–1097.
- Harnard, S. (1990). The symbol grounding problem. *Physica D* 42, 335–346.
- Harvey, A. C. (1990). *Forecasting, Structural Time Series Models, and the Kalman Filter*. Cambridge University Press.

- Hastie, T., S. Rosset, R. Tibshirani, and J. Zhu (2004). The entire regularization path for the support vector machine. *J. of Machine Learning Research* 5, 1391–1415.
- Hastie, T. and R. Tibshirani (1990). *Generalized additive models*. Chapman and Hall.
- Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer.
- Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. Springer. 2nd edition.
- Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika* 57(1), 97–109.
- Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation*. Prentice Hall. 2nd Edition.
- Haykin, S. (Ed.) (2001). *Kalman Filtering and Neural Networks*. Wiley.
- Hazan, T. and A. Shashua (2008). Convergent message-passing algorithms for inference over general graphs with convex free energy. In *UAI*.
- Hazan, T. and A. Shashua (2010). Norm-product belief propagation: primal-dual message passing for approximate inference. *IEEE Trans. on Info. Theory* 56(12), 6294–6316.
- He, Y.-B. and Z. Geng (2009). Active learning of causal networks with intervention experiments and optimal designs. *J. of Machine Learning Research* 10, 2523–2547.
- Heaton, M. and J. Scott (2009). Bayesian computation and the linear model. Technical report, Duke.
- Heckerman, D., D. Chickering, C. Meek, R. Rounthwaite, and C. Kadie (2000). Dependency networks for knowledge estimation, collaborative filtering, and data visualization. *J. of Machine Learning Research* 1, 49–75.
- Heckerman, D., D. Geiger, and M. Chickering (1995). Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning* 20(3), 197–243.
- Heckerman, D., C. Meek, and G. Cooper (1997, February). A Bayesian approach to causal discovery. Technical Report MSR-TR-97-05, Microsoft Research.
- Heckerman, D., C. Meek, and D. Koller (2004). Probabilistic models for relational data. Technical Report MSR-TR-2004-30, Microsoft Research.
- Heller, K. and Z. Ghahramani (2005). Bayesian Hierarchical Clustering. In *Intl. Conf. on Machine Learning*.
- Henrion, M. (1988). Propagation of uncertainty by logic sampling in Bayes' networks. In *UAI*, pp. 149–164.
- Herbrich, R., T. Minka, and T. Graepel (2007). TrueSkill: A Bayesian skill rating system. In *NIPS*.
- Hertz, J., A. Krogh, and R. G. Palmer (1991). *An Introduction to the Theory of Neural Computation*. Addison-Wesley.
- Hillar, C., J. Sohl-Dickstein, and K. Koepsell (2012, April). Efficient and optimal binary hopfield associative memory storage using minimum probability flow. Technical report.
- Hinton, G. (1999). Products of experts. In *Proc. 9th Intl. Conf. on Artif. Neural Networks (ICANN)*, Volume 1, pp. 1–6.
- Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation* 14, 1771–1800.
- Hinton, G. (2010). A Practical Guide to Training Restricted Boltzmann Machines. Technical report, U. Toronto.
- Hinton, G. and D. V. Camp (1993). Keeping neural networks simple by minimizing the description length of the weights. In *in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*, pp. 5–13. ACM Press.
- Hinton, G., S. Osindero, and Y. Teh (2006). A fast learning algorithm for deep belief nets. *Neural Computation* 18, 1527–1554.
- Hinton, G. and R. Salakhutdinov (2006, July). Reducing the dimensionality of data with neural networks. *Science* 313(5786), 504–507.
- Hinton, G. E., P. Dayan, and M. Revow (1997). Modeling the manifolds of images of handwritten digits. *IEEE Trans. on Neural Networks* 8, 65–74.
- Hinton, G. E. and Y. Teh (2001). Discovering multiple constraints that are frequently approximately satisfied. In *UAI*.
- Hjort, N., C. Holmes, P. Muller, and S. Walker (Eds.) (2010). *Bayesian Nonparametrics*. Cambridge.
- Hoeffling, H. (2010). A Path Algorithm for the Fused Lasso Signal Approximator. Technical report, Stanford.
- Hoeffling, H. and R. Tibshirani (2009). Estimation of Sparse Binary Pairwise Markov Networks using Pseudo-likelihoods. *J. of Machine Learning Research* 10.
- Hoeting, J., D. Madigan, A. Raftery, and C. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science* 4(4).
- Hoff, P. D. (2009, July). *A First Course in Bayesian Statistical Methods*. Springer.
- Hoffman, M., D. Blei, and F. Bach (2010). Online learning for latent dirichlet allocation. In *NIPS*.
- Hoffman, M. and A. Gelman (2011). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. Technical report, Columbia U.
- Hofmann, T. (1999). Probabilistic latent semantic indexing. *Research and Development in Information Retrieval*, 50–57.
- Holmes, C. and L. Held (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* 1(1), 145–168.
- Honkela, A. and H. Valpola (2004). Variational Learning and Bits-Back Coding: An Information-Theoretic View to Bayesian Learning. *IEEE Trans. on Neural Networks* 15(4).
- Honkela, A., H. Valpola, and J. Karhunen (2003). Accelerating Cyclic Update Algorithms for Parameter Estimation by Pattern Searches. *Neural Processing Letters* 17, 191–203.

- Hopfield, J. J. (1982, April). Neural networks and physical systems with emergent collective computational abilities. *Proc. of the National Academy of Science, USA* 79(8), 2554–2558.
- Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks* 4(2), 251–257.
- Horvitz, E., J. Apacible, R. Sarin, and L. Liao (2005). Prediction, Expectation, and Surprise: Methods, Designs, and Study of a Deployed Traffic Forecasting Service. In *UAI*.
- Howard, R. and J. Matheson (1981). Influence diagrams. In R. Howard and J. Matheson (Eds.), *Readings on the Principles and Applications of Decision Analysis, volume II*. Strategic Decisions Group.
- Hoyer, P. (2004). Non-negative matrix factorization with sparseness constraints. *J. of Machine Learning Research* 5, 1457–1469.
- Hsu, C.-W., C.-C. Chang, and C.-J. Lin (2009). A practical guide to support vector classification. Technical report, Dept. Comp. Sci., National Taiwan University.
- Hu, D., L. van der Maaten, Y. Cho, L. Saul, and S. Lerner (2010). Latent Variable Models for Predicting File Dependencies in Large-Scale Software Development. In *NIPS*.
- Hu, M., C. Ingram, M. Sirski, C. Pal, S. Swamy, and C. Patten (2000). A Hierarchical HMM Implementation for Vertebrate Gene Splice Site Prediction. Technical report, Dept. Computer Science, Univ. Waterloo.
- Huang, J., Q. Morris, and B. Frey (2007). Bayesian inference of MicroRNA targets from sequence and expression data. *J. Comp. Bio.*
- Hubel, D. and T. Wiesel (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J. Physiology* 160, 106–154.
- Huber, P. (1964). Robust estimation of a location parameter. *Annals of Statistics* 53, 73–101.
- Hubert, L. and P. Arabie (1985). Comparing partitions. *J. of Classification* 2, 193–218.
- Hunter, D. and R. Li (2005). Variable selection using MM algorithms. *Annals of Statistics* 33, 1617–1642.
- Hunter, D. R. and K. Lange (2004). A Tutorial on MM Algorithms. *The American Statistician* 58, 30–37.
- Hyafil, L. and R. Rivest (1976). Constructing Optimal Binary Decision Trees is NP-complete. *Information Processing Letters* 5(1), 15–17.
- Hyvarinen, A., J. Hurri, and P. Hoyer (2009). *Natural Image Statistics: a probabilistic approach to early computational vision*. Springer.
- Hyvarinen, A. and E. Oja (2000). Independent component analysis: algorithms and applications. *Neural Networks* 13, 411–430.
- Ilin, A. and T. Raiko (2010). Practical Approaches to Principal Component Analysis in the Presence of Missing Values. *J. of Machine Learning Research* 11, 1957–2000.
- Insua, D. R. and F. Ruggeri (Eds.) (2000). *Robust Bayesian Analysis*. Springer.
- Isard, M. (2003). PAMPAS: Real-Valued Graphical Models for Computer Vision. In *CVPR*, Volume 1, pp. 613.
- Isard, M. and A. Blake (1998). CONDENSATION - conditional density propagation for visual tracking. *Intl. J. of Computer Vision* 29(1), 5–18.
- Jaakkola, T. (2001). Tutorial on variational approximation methods. In M. Oppen and D. Saad (Eds.), *Advanced mean field methods*. MIT Press.
- Jaakkola, T. and D. Haussler (1998). Exploiting generative models in discriminative classifiers. In *NIPS*, pp. 487–493.
- Jaakkola, T. and M. Jordan (1996a). Computing upper and lower bounds on likelihoods in intractable networks. In *UAI*.
- Jaakkola, T. and M. Jordan (1996b). A variational approach to Bayesian logistic regression problems and their extensions. In *AI + Statistics*.
- Jaakkola, T. S. and M. I. Jordan (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* 10, 25–37.
- Jacob, L., F. Bach, and J.-P. Vert (2008). Clustered Multi-Task Learning: a Convex Formulation. In *NIPS*.
- Jain, A. and R. Dubes (1988). *Algorithms for Clustering Data*. Prentice Hall.
- James, G. and T. Hastie (1998). The error coding method and PICTS. *J. of Computational and Graphical Statistics* 7(3), 377–387.
- Japkowicz, N., S. Hanson, and M. Gluck (2000). Nonlinear autoassociation is not equivalent to PCA. *Neural Computation* 12, 531–545.
- Jaynes, E. T. (2003). *Probability theory: the logic of science*. Cambridge university press.
- Jebara, T., R. Kondor, and A. Howard (2004). Probability product kernels. *J. of Machine Learning Research* 5, 819–844.
- Jeffreys, H. (1961). *Theory of Probability*. Oxford.
- Jelinek, F. (1997). *Statistical methods for speech recognition*. MIT Press.
- Jensen, C. S., A. Kong, and U. Kjaerulff (1995). Blocking-gibbs sampling in very large probabilistic expert systems. *Intl. J. Human-Computer Studies*, 647–666.
- Jermyn, I. (2005). Invariant bayesian estimation on manifolds. *Annals of Statistics* 33(2), 583–605.
- Jerrum, M. and A. Sinclair (1993). Polynomial-time approximation algorithms for the Ising model. *SIAM J. on Computing* 22, 1087–1116.
- Jerrum, M. and A. Sinclair (1996). The markov chain monte carlo method: an approach to approximate counting and integration. In D. S. Hochbaum (Ed.), *Approximation Algorithms for NP-hard problems*. PWS Publishing.
- Jerrum, M., A. Sinclair, and E. Vigoda (2004). A polynomial-time approximation algorithm for the permanent of a matrix with non-negative entries. *Journal of the ACM*, 671–697.
- Ji, S., D. Dunson, and L. Carin (2009). Multi-task compressive sensing. *IEEE Trans. Signal Processing* 57(1).

- Ji, S., L. Tang, S. Yu, and J. Ye (2010). A shared-subspace learning framework for multi-label classification. *ACM Trans. on Knowledge Discovery from Data* 4(2).
- Jirousek, R. and S. Preucil (1995). On the effective implementation of the iterative proportional fitting procedure. *Computational Statistics & Data Analysis* 19, 177–189.
- Joachims, T. (2006). Training Linear SVMs in Linear Time. In *Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining*.
- Joachims, T., T. Finley, and C.-N. Yu (2009). Cutting-Plane Training of Structural SVMs. *Machine Learning* 77(1), 27–59.
- Johnson, J. K., D. M. Malioutov, and A. S. Willsky (2006). Walk-sum interpretation and analysis of gaussian belief propagation. In *NIPS*, pp. 579–586.
- Johnson, M. (2005). Capacity and complexity of HMM duration modeling techniques. *Signal Processing Letters* 12(5), 407–410.
- Johnson, N. (2009). A study of the NIPS feature selection challenge. Technical report, Stanford.
- Johnson, V. and J. Albert (1999). *Ordinal data modeling*. Springer.
- Jones, B., A. Dobra, C. Carvalho, C. Hans, C. Carter, and M. West (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science* 20, 388–400.
- Jordan, M. I. (2007). An introduction to probabilistic graphical models. In preparation.
- Jordan, M. I. (2011). The era of big data. In *ISBA Bulletin*, Volume 18, pp. 1–3.
- Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1998). An introduction to variational methods for graphical models. In M. Jordan (Ed.), *Learning in Graphical Models*. MIT Press.
- Jordan, M. I. and R. A. Jacobs (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6, 181–214.
- Journee, M., Y. Nesterov, P. Richtarik, and R. Sepulchre (2010). Generalized power method for sparse principal components analysis. *J. of Machine Learning Research* 11, 517–553.
- Julier, S. and J. Uhlmann (1997). A new extension of the Kalman filter to nonlinear systems. In *Proc. of AeroSense: The 11th Intl. Symp. on Aerospace/Defence Sensing, Simulation and Controls*.
- Jurafsky, D. and J. H. Martin (2000). *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall.
- Jurafsky, D. and J. H. Martin (2008). *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall. 2nd edition.
- Kaariainen, M. and J. Langford (2005). A Comparison of Tight Generalization Bounds. In *Intl. Conf. on Machine Learning*.
- Kaelbling, L., M. Littman, and A. Moore (1996). Reinforcement learning: A survey. *J. of AI Research* 4, 237–285.
- Kaelbling, L. P., M. Littman, and A. Cassandra (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence* 101.
- Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23(3).
- Kakade, S., Y. W. Teh, and S. Roweis (2002). An alternate objective function for markovian fields. In *Intl. Conf. on Machine Learning*.
- Kanazawa, K., D. Koller, and S. Russell (1995). Stochastic simulation algorithms for dynamic probabilistic networks. In *UAI*.
- Kandel, E., J. Schwartz, and T. Jessell (2000). *Principles of Neural Science*. McGraw-Hill.
- Kappen, H. and F. Rodriguez (1998). Boltzmann machine learning using mean field theory and linear response correction. In *NIPS*.
- Karhunen, J. and J. Joutsensalo (1995). Generalizations of principal component analysis, optimization problems, and neural networks. *Neural Networks* 8(4), 549–562.
- Kass, R. and L. Wasserman (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *J. of the Am. Stat. Assoc.* 90(431), 928–934.
- Katayama, T. (2005). *Subspace Methods for Systems Identification*. Springer Verlag.
- Kaufman, L. and P. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley.
- Kawakatsu, H. and A. Largey (2009). EM algorithms for ordered probit models with endogenous regressors. *The Econometrics Journal* 12(1), 164–186.
- Kearns, M. J. and U. V. Vazirani (1994). *An Introduction to Computational Learning Theory*. MIT Press.
- Kelley, J. E. (1960). The cutting-plane method for solving convex programs. *J. of the Soc. for Industrial and Applied Math.* 8, 703–712.
- Kemp, C., J. Tenenbaum, S. Niyogi, and T. Griffiths (2010). A probabilistic model of theory formation. *Cognition* 114, 165–196.
- Kemp, C., J. Tenenbaum, T. Y. T. Griffiths and, and N. Ueda (2006). Learning systems of concepts with an infinite relational model. In *AAAI*.
- Kersting, K., S. Natarajan, and D. Poole (2011). *Statistical Relational AI: Logic, Probability and Computation*. Technical report, UBC.
- Khan, M. E., B. Marlin, G. Bouchard, and K. P. Murphy (2010). Variational bounds for mixed-data factor analysis. In *NIPS*.
- Khan, Z., T. Balch, and F. Dellaert (2006). MCMC Data Association and Sparse Factorization Updating for Real Time Multitarget Tracking with Merged and Multiple Measurements. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 28(12).
- Kirkpatrick, S., C. G. Jr., and M. Vecchi (1983). Optimization by simulated annealing. *Science* 220, 671–680.

- Kitagawa, G. (2004). The two-filter formula for smoothing and an implementation of the Gaussian-sum smoother. *Annals of the Institute of Statistical Mathematics* 46(4), 605–623.
- Kjaerulff, U. (1990). Triangulation of graphs – algorithms giving small total state space. Technical Report R-90-09, Dept. of Math. and Comp. Sci., Aalborg Univ., Denmark.
- Kjaerulff, U. and A. Madsen (2008). *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*. Springer.
- Klaassen, C. and J. A. Wellner (1997). Efficient estimation in the bivariate normal copula model: Normal margins are least favorable. *Bernoulli* 3(1), 55–77.
- Klami, A. and S. Kaski (2008). Probabilistic approach to detecting dependencies between data sets. *Neurocomputing* 72, 39–46.
- Klami, A., S. Virtanen, and S. Kaski (2010). Bayesian exponential family projections for coupled data sources. In *UAI*.
- Kleiner, A., A. Talwalkar, P. Sarkar, and M. I. Jordan (2011). A scalable bootstrap for massive data. Technical report, UC Berkeley.
- Kneser, R. and H. Ney (1995). Improved backing-off for n-gram language modeling. In *Intl. Conf. on Acoustics, Speech and Signal Proc.*, Volume 1, pp. 181–184.
- Ko, J. and D. Fox (2009). GP-BayesFilters: Bayesian Filtering Using Gaussian Process Prediction and Observation Models. *Autonomous Robots Journal*.
- Kohn, R., M. Smith, and D. Chan (2001). Nonparametric regression using linear combinations of basis functions. *Statistical Computing* 11, 313–322.
- Koivisto, M. (2006). Advances in exact Bayesian structure discovery in Bayesian networks. In *UAI*.
- Koivisto, M. and K. Sood (2004). Exact Bayesian structure discovery in Bayesian networks. *J. of Machine Learning Research* 5, 549–573.
- Koller, D. and N. Friedman (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press.
- Koller, D. and U. Lerner (2001). Sampling in Factored Dynamic Systems. In A. Doucet, N. de Freitas, and N. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*. Springer.
- Kolmogorov, V. (2006, October). Convergent Tree-reweighted Message Passing for Energy Minimization. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 28(10), 1568–1583.
- Kolmogorov, V. and M. Wainwright (2005). On optimality properties of tree-reweighted message passing. In *UAI*, pp. 316–322.
- Kolmogorov, V. and R. Zabini (2004). What energy functions can be minimized via graph cuts? *IEEE Trans. on Pattern Analysis and Machine Intelligence* 26(2), 147–159.
- Komodakis, N., N. Paragios, and G. Tziritas (2011). MRF Energy Minimization and Beyond via Dual Decomposition. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 33(3), 531–552.
- Koo, T., A. M. Rush, M. Collins, T. Jaakkola, and D. Sontag (2010). Dual Decomposition for Parsing with Non-Projective Head Automata. In *Proc. EMNLP*, pp. 1288–1298.
- Koren, Y. (2009a). The bellkor solution to the netflix grand prize. Technical report, Yahoo! Research.
- Koren, Y. (2009b). Collaborative filtering with temporal dynamics. In *Proc. of the Intl Conf. on Knowledge Discovery and Data Mining*.
- Koren, Y., R. Bell, and C. Volinsky (2009). Matrix factorization techniques for recommender systems. *IEEE Computer* 42(8), 30–37.
- Krishnapuram, B., L. Carin, M. Figueiredo, and A. Hartemink (2005). Learning sparse bayesian classifiers: multi-class formulation, fast algorithms, and generalization bounds. *IEEE Transaction on Pattern Analysis and Machine Intelligence*.
- Krizhevsky, A. and G. Hinton (2010). Using Very Deep Autoencoders for Content-Based Image Retrieval. Submitted.
- Kschischang, F., B. Frey, and H.-A. Loeliger (2001, February). Factor graphs and the sum-product algorithm. *IEEE Trans Info. Theory*.
- Kuan, P., G. Pan, J. A. Thomson, R. Stewart, and S. Keles (2009). A hierarchical semi-Markov model for detecting enrichment with application to ChIP-Seq experiments. Technical report, U. Wisconsin.
- Kulesza, A. and B. Taskar (2011). Learning Determinantal Point Processes. In *UAI*.
- Kumar, N. and A. Andreo (1998). Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Communication* 26, 283–297.
- Kumar, S. and M. Hebert (2003). Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Intl. Conf. on Computer Vision*.
- Kuo, L. and B. Mallick (1998). Variable selection for regression models. *Sankhya Series B* 60, 65–81.
- Kurihara, K., M. Welling, and N. Vlassis (2006). Accelerated variational DP mixture models. In *NIPS*.
- Kushner, H. and G. Yin (2003). *Stochastic approximation and recursive algorithms and applications*. Springer.
- Kuss and C. Rasmussen (2005). Assessing approximate inference for binary gaussian process classification. *J. of Machine Learning Research* 6, 1679–1704.
- Kwon, J. and K. Murphy (2000). Modeling freeway traffic with coupled HMMs. Technical report, Univ. California, Berkeley.
- Kyung, M., J. Gill, M. Ghosh, and G. Casella (2010). Penalized Regression, Standard Errors and Bayesian Lasso. *Bayesian Analysis* 5(2), 369–412.
- Lacoste-Julien, S., F. Huszar, and Z. Ghahramani (2011). Approximate inference for the loss-calibrated Bayesian. In *AI/Statistics*.
- Lacoste-Julien, S., F. Sha, and M. I. Jordan (2009). DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*.

- Lafferty, J., A. McCallum, and F. Pereira (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Intl. Conf. on Machine Learning*.
- Lange, K., R. Little, and J. Taylor (1989). Robust statistical modeling using the t distribution. *J. of the Am. Stat. Assoc.* 84(408), 881–896.
- Langville, A. and C. Meyer (2006). Updating Markov chains with an eye on Google's PageRank. *SIAM J. on Matrix Analysis and Applications* 27(4), 968–987.
- Larranaga, P., C. M. H. Kuijpers, M. Poza, and R. H. Murga (1997). Decomposing bayesian networks: triangulation of the moral graph with genetic algorithms. *Statistics and Computing (UK)* 7(1), 19–34.
- Lashkari, D. and P. Golland (2007). Convex clustering with exemplar-based models. In *NIPS*.
- Lasserre, J., C. Bishop, and T. Minka (2006). Principled hybrids of generative and discriminative models. In *CVPR*.
- Lau, J. and P. Green (2006). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* 12, 351–357.
- Lauritzen, S. (1996). *Graphical Models*. OUP.
- Lauritzen, S. (2000). Causal inference from graphical models. In D. R. C. O. E. Barndoff-Nielsen and C. Klueppelberg (Eds.), *Complex stochastic systems*. Chapman and Hall.
- Lauritzen, S. and D. Nilsson (2001). Representing and solving decision problems with limited information. *Management Science* 47, 1238–1251.
- Lauritzen, S. L. (1992, December). Propagation of probabilities, means and variances in mixed graphical association models. *J. of the Am. Stat. Assoc.* 87(420), 1098–1108.
- Lauritzen, S. L. (1995). The EM algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis* 19, 191–201.
- Lauritzen, S. L. and D. J. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their applications to expert systems. *J. R. Stat. Soc. B* 50, 127–224.
- Law, E., B. Settles, and T. Mitchell (2010). Learning to tag from open vocabulary labels. In *Proc. European Conf. on Machine Learning*.
- Law, M., M. Figueiredo, and A. Jain (2004). Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 26(4).
- Lawrence, N. D. (2005). Probabilistic non-linear principal component analysis with gaussian process latent variable models. *J. of Machine Learning Research* 6, 1783–1816.
- Lawrence, N. D. (2012). A unifying probabilistic perspective for spectral dimensionality reduction: insights and new models. *J. of Machine Learning Research* 13, 1609–1638.
- Learned-Miller, E. (2004). Hyperspacing and the estimation of information theoretic quantities. Technical Report 04-104, U. Mass. Amherst Comp. Sci. Dept.
- LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel (1989, Winter). Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1(4), 541–551.
- LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998, November). Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11), 2278–2324.
- LeCun, Y., S. Chopra, R. Hadsell, F.-J. Huang, and M.-A. Ranzato (2006). A tutorial on energy-based learning. In B. et al. (Ed.), *Predicting Structured Outputs*. MIT press.
- Ledoit, O. and M. Wolf (2004a). Honey, I Shrunk the Sample Covariance Matrix. *J. of Portfolio Management* 31(1).
- Ledoit, O. and M. Wolf (2004b). A well-conditioned estimator for large-dimensional covariance matrices. *J. of Multivariate Analysis* 88(2), 365–411.
- Lee, A., F. Caron, A. Doucet, and C. Holmes (2010). A hierarchical bayesian framework for constructing sparsity-inducing priors. Technical report, U. Oxford.
- Lee, A., F. Caron, A. Doucet, and C. Holmes (2011). Bayesian Sparsity-Path-Analysis of Genetic Association Signal using Generalized t Prior. Technical report, U. Oxford.
- Lee, D. and S. Seung (2001). Algorithms for non-negative matrix factorization. In *NIPS*.
- Lee, H., R. Grosse, R. Ranganath, and A. Ng (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Intl. Conf. on Machine Learning*.
- Lee, H., Y. Largman, P. Pham, and A. Ng (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *NIPS*.
- Lee, S.-I., V. Ganapathi, and D. Koller (2006). Efficient structure learning of Markov networks using L1-regularization. In *NIPS*.
- Lee, T. S. and D. Mumford (2003). Hierarchical Bayesian inference in the visual cortex. *J. of Optical Society of America A* 20(7), 1434–1448.
- Lenk, P., W. S. DeSarbo, P. Green, and M. Young (1996). Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs. *Marketing Science* 15(2), 173–191.
- Lenkoski, A. and A. Dobra (2008). Bayesian structural learning and estimation in Gaussian graphical models. Technical Report 545, Department of Statistics, University of Washington.
- Lepar, V. and P. P. Shenoy (1998). A Comparison of Lauritzen-Spiegelhalter, Hugin and Shenoy-Shafer Architectures for Computing Marginals of Probability Distributions. In G. Cooper and S. Moral (Eds.), *UAI*, pp. 328–337. Morgan Kaufmann.
- Lerner, U. and R. Parr (2001). Inference in hybrid networks: Theoretical limits and practical algorithms. In *UAI*.

- Leslie, C., E. Eskin, A. Cohen, J. Weston, and W. Noble (2003). Mismatch string kernels for discriminative protein classification. *Bioinformatics* 1, 1–10.
- Levy, S. (2011). *In The Plex: How Google Thinks, Works, and Shapes Our Lives*. Simon & Schuster.
- Li, L., W. Chu, J. Langford, and X. Wang (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *WSDM*.
- Liang, F., S. Mukherjee, and M. West (2007). Understanding the use of unlabelled data in predictive modelling. *Statistical Science* 22, 189–205.
- Liang, F., R. Paulo, G. Molina, M. Clyde, and J. Berger (2008). Mixtures of g-priors for Bayesian Variable Selection. *J. of the Am. Stat. Assoc.* 103(481), 410–423.
- Liang, P. and M. I. Jordan (2008). An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *International Conference on Machine Learning (ICML)*.
- Liang, P. and D. Klein. Online EM for Unsupervised Models. In *Proc. NAACL Conference*.
- Liao, L., D. J. Patterson, D. Fox, and H. Kautz (2007). Learning and Inferring Transportation Routines. *Artificial Intelligence* 17(5), 311–331.
- Lindley, D. (1982). Scoring rules and the inevitability of probability. *ISI Review* 50, 1–26.
- Lindley, D. V. (1972). *Bayesian Statistics: A Review*. SIAM.
- Lindley, D. V. and L. D. Phillips (1976). Inference for a Bernoulli Process (A Bayesian View). *The American Statistician* 30(3), 112–119.
- Lindsay, B. (1988). Composite likelihood methods. *Contemporary Mathematics* 80(1), 221–239.
- Lipton, R. J. and R. E. Tarjan (1979). A separator theorem for planar graphs. *SIAM Journal of Applied Math* 36, 177–189.
- Little, R. J. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. New York: Wiley and Son.
- Liu, C. and D. Rubin (1995). ML Estimation of the T distribution using EM and its extensions, ECM and ECME. *Statistica Sinica* 5, 19–39.
- Liu, H., J. Lafferty, and L. Wasserman (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *J. of Machine Learning Research* 10, 2295–2328.
- Liu, J. (2001). *Monte Carlo Strategies in Scientific Computation*. Springer.
- Liu, J. S., W. H. Wong, and A. Kong (1994). Covariance structure of the gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* 81(1), 27–40.
- Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval* 3(3), 225–331.
- Lizotte, D. (2008). *Practical Bayesian optimization*. Ph.D. thesis, U. Alberta.
- Ljung, L. (1987). *System Identification: Theory for the User*. Prentice Hall.
- Lo, C. H. (2009). *Statistical methods for high throughput genomics*. Ph.D. thesis, UBC.
- Lo, K., F. Hahne, R. Brinkman, R. Ryan, and R. Gottardo (2009, May). flowclust: a bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics* 10, 145+.
- Lopes, H. and M. West (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* 14, 41–67.
- Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision ICCV, Corfu*, pp. 1150–1157.
- Luce, R. (1959). *Individual choice behavior: A theoretical analysis*. Wiley.
- Lunn, D., N. Best, and J. Whittaker (2009). Generic reversible jump MCMC using graphical models. *Statistics and Computing* 19(4), 395–408.
- Lunn, D., A. Thomas, N. Best, and D. Spiegelhalter (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 10, 325–337.
- Ma, H., H. Yang, M. Lyu, and I. King (2008). SoRec: Social recommendation using probabilistic matrix factorization. In *Proc. of 17th Conf. on Information and Knowledge Management*.
- Ma, S., C. Ji, and J. Farmer (1997). An efficient EM-based training algorithm for feedforward neural networks. *Neural Networks* 10(2), 243–256.
- Maathuis, M., D. Colombo, M. Kalisch, and P. Bajhlmann (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods* 7, 247–248.
- Maathuis, M., M. Kalisch, and P. Bajhlmann (2009). Estimating high-dimensional intervention effects from observational data. *Annals of Statistics* 37, 3133–3164.
- MacKay, D. (1992). Bayesian interpolation. *Neural Computation* 4, 415–447.
- MacKay, D. (1995a). Developments in probabilistic modeling with neural networks — ensemble learning. In *Proc. 3rd Ann. Symp. Neural Networks*.
- MacKay, D. (1995b). Probable networks and plausible predictions — a review of practical Bayesian methods for supervised neural networks. *Network*.
- MacKay, D. (1997). Ensemble learning for Hidden Markov Models. Technical report, U. Cambridge.
- MacKay, D. (1999). Comparison of approximate methods for handling hyperparameters. *Neural Computation* 11(5), 1035–1068.
- MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.
- Macnaughton-Smith, P., W. T. Williams, M. B. Dale, and G. Mockett (1964). Dissimilarity analysis: a new technique of hierarchical sub-division. *Nature* 202, 1034 – 1035.

- Madeira, S. C. and A. L. Oliveira (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 1(1), 24–45.
- Madigan, D. and A. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. of the Am. Stat. Assoc.* 89, 1535–1546.
- Madsen, R., D. Kauchak, and C. Elkan (2005). Modeling word burstiness using the Dirichlet distribution. In *Intl. Conf. on Machine Learning*.
- Mairal, J., F. Bach, J. Ponce, and G. Sapiro (2010). Online learning for matrix factorization and sparse coding. *J. of Machine Learning Research* 11, 19–60.
- Mairal, J., M. Elad, and G. Sapiro (2008). Sparse representation for color image restoration. *IEEE Trans. on Image Processing* 17(1), 53–69.
- Malioutov, D., J. Johnson, and A. Wilksy (2006). Walk-sums and belief propagation in gaussian graphical models. *J. of Machine Learning Research* 7, 2003–2030.
- Mallat, S., G. Davis, and Z. Zhang (1994, July). Adaptive time-frequency decompositions. *SPIE Journal of Optical Engineering* 33, 2183–2919.
- Mallat, S. and Z. Zhang (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing* 41(12), 3397–3415.
- Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *Proc. Sixth Conference on Natural Language Learning (CoNLL-2002)*, pp. 49–55.
- Manning, C., P. Raghavan, and H. Schuetze (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Manning, C. and H. Schuetze (1999). *Foundations of statistical natural language processing*. MIT Press.
- Mansinghka, V., D. Roy, R. Rifkin, and J. Tenenbaum (2007). Aclass: An online algorithm for generative classification. In *AI/Statistics*.
- Mansinghka, V., P. Shafto, E. Jonas, C. Petschulat, and J. Tenenbaum (2011). Cross-Categorization: A Nonparametric Bayesian Method for Modeling Heterogeneous, High Dimensional Data. Technical report, MIT.
- Margolin, A., I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, and R. F. abd A. Califano (2006). ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* 7.
- Marin, J.-M. and C. Robert (2007). *Bayesian Core: a practical approach to computational Bayesian statistics*. Springer.
- Marks, T. K. and J. R. Movellan (2001). Diffusion networks, products of experts, and factor analysis. Technical report, University of California San Diego.
- Marlin, B. (2003). Modeling user rating profiles for collaborative filtering. In *NIPS*.
- Marlin, B. (2008). *Missing Data Problems in Machine Learning*. Ph.D. thesis, U. Toronto.
- Marlin, B., E. Khan, and K. Murphy (2011). Piecewise Bounds for Estimating Bernoulli-Logistic Latent Gaussian Models. In *Intl. Conf. on Machine Learning*.
- Marlin, B. and R. Zemel (2009). Collaborative prediction and ranking with non-random missing data. In *Proc. of the 3rd ACM Conference on Recommender Systems*.
- Marlin, B. M., K. Swersky, B. Chen, and N. de Freitas (2010). Inductive principles for restricted boltzmann machine learning. In *AI/Statistics*.
- Marroquin, J., S. Mitter, and T. Poggio (1987). Probabilistic solution of ill-posed problems in computational vision. *J. of the Am. Stat. Assoc.* 82(297), 76–89.
- Martens, J. (2010). Deep learning via hessian-free optimization. In *Intl. Conf. on Machine Learning*.
- Maruyama, Y. and E. George (2008). A g-prior extension for $p > n$. Technical report, U. Tokyo.
- Mason, L., J. Baxter, P. Bartlett, and M. Frean (2000). Boosting algorithms as gradient descent. In *NIPS*, Volume 12, pp. 512–518.
- Matthews, R. (1998). *Bayesian Critique of Statistics in Health: The Great Health Hoax*.
- Maybeck, P. (1979). *Stochastic models, estimation, and control*. Academic Press.
- Mazumder, R. and T. Hastie (2012). The Graphical Lasso: New Insights and Alternatives. Technical report.
- McAuliffe, J., D. Blei, and M. Jordan (2006). Nonparametric empirical bayes for the dirichlet process mixture model. *Statistics and Computing* 16(1), 5–14.
- McCallum, A. (2003). Efficiently inducing features of conditional random fields. In *UAI*.
- McCallum, A., D. Freitag, and F. Pereira (2000). Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Intl. Conf. on Machine Learning*.
- McCallum, A. and K. Nigam (1998). A comparison of event models for naive Bayes text classification. In *AAAI/ICML workshop on Learning for Text Categorization*.
- McCray, A. (2003). An upper level ontology for the biomedical domain. *Comparative and Functional Genomics* 4, 80–84.
- McCullagh, P. and J. Nelder (1989). *Generalized linear models*. Chapman and Hall. 2nd edition.
- McCulloch, W. and W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5, 115–137.
- McDonald, J. and W. Newey (1988). Partially Adaptive Estimation of Regression Models via the Generalized t Distribution. *Econometric Theory* 4(3), 428–445.
- McEliece, R. J., D. J. C. MacKay, and J. F. Cheng (1998). Turbo decoding as an instance of Pearl's 'belief propagation' algorithm. *IEEE J. on Selected Areas in Comm.* 16(2), 140–152.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics*, pp. 105–142. Academic Press.

- McGrayne, S. B. (2011). *The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*. Yale University Press.
- McKay, B. D., F. E. Oggier, G. F. Royle, N. J. A. Sloane, I. M. Wanless, and H. S. Wilf (2004). Acyclic digraphs and eigenvalues of (0,1)-matrices. *J. Integer Sequences* 7(04.3.3).
- McKay, D. and L. C. B. Peto (1995). A hierarchical dirichlet language model. *Natural Language Engineering* 1(3), 289–307.
- McLachlan, G. J. and T. Krishnan (1997). *The EM Algorithm and Extensions*. Wiley.
- Meek, C. and D. Heckerman (1997). Structure and parameter learning for causal independence and causal interaction models. In *UAI*, pp. 366–375.
- Meek, C., B. Thiesson, and D. Heckerman (2002). Staged mixture modelling and boosting. In *UAI*, San Francisco, CA, pp. 335–343. Morgan Kaufmann.
- Meila, M. (2001). A random walks view of spectral segmentation. In *AI/Statistics*.
- Meila, M. (2005). Comparing clusterings: an axiomatic view. In *Intl. Conf. on Machine Learning*.
- Meila, M. and T. Jaakkola (2006). Tractable Bayesian learning of tree belief networks. *Statistics and Computing* 16, 77–92.
- Meila, M. and M. I. Jordan (2000). Learning with mixtures of trees. *J. of Machine Learning Research* 1, 1–48.
- Meinshausen, N. (2005). A note on the lasso for gaussian graphical model selection. Technical report, ETH Seminar fur Statistik.
- Meinshausen, N. and P. Bühlmann (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34, 1436–1462.
- Meinshausen, N. and P. Bühlmann (2010). Stability selection. *J. of Royal Stat. Soc. Series B* 72, 417–473.
- Meltzer, T., C. Yanover, and Y. Weiss (2005). Globally optimal solutions for energy minimization in stereo vision using reweighted belief propagation. In *ICCV*, pp. 428–435.
- Meng, X. L. and D. van Dyk (1997). The EM algorithm — an old folk song sung to a fast new tune (with Discussion). *J. Royal Stat. Soc. B* 59, 511–567.
- Mesot, B. and D. Barber (2009). A Simple Alternative Derivation of the Expectation Correction Algorithm. *IEEE Signal Processing Letters* 16(1), 121–124.
- Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *J. of Chemical Physics* 21, 1087–1092.
- Metz, C. (2010). Google behavioral ad targeter is a Smart Ass. *The Register*.
- Miller, A. (2002). *Subset selection in regression*. Chapman and Hall. 2nd edition.
- Mimno, D. and A. McCallum (2008). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *UAI*.
- Minka, T. (1999). Pathologies of orthodox statistics. Technical report, MIT Media Lab.
- Minka, T. (2000a). Automatic choice of dimensionality for PCA. Technical report, MIT.
- Minka, T. (2000b). Bayesian linear regression. Technical report, MIT.
- Minka, T. (2000c). Bayesian model averaging is not model combination. Technical report, MIT Media Lab.
- Minka, T. (2000d). Empirical risk minimization is an incomplete inductive principle. Technical report, MIT.
- Minka, T. (2000e). Estimating a Dirichlet distribution. Technical report, MIT.
- Minka, T. (2000f). Inferring a Gaussian distribution. Technical report, MIT.
- Minka, T. (2001a). Bayesian inference of a uniform distribution. Technical report, MIT.
- Minka, T. (2001b). Empirical Risk Minimization is an incomplete inductive principle. Technical report, MIT.
- Minka, T. (2001c). Expectation propagation for approximate Bayesian inference. In *UAI*.
- Minka, T. (2001d). *A family of algorithms for approximate Bayesian inference*. Ph.D. thesis, MIT.
- Minka, T. (2001e). Statistical approaches to learning and discovery 10-602: Homework assignment 2, question 5. Technical report, CMU.
- Minka, T. (2003). A comparison of numerical optimizers for logistic regression. Technical report, MSR.
- Minka, T. (2005). Divergence measures and message passing. Technical report, MSR Cambridge.
- Minka, T. and Y. Qi (2003). Tree-structured approximations by expectation propagation. In *NIPS*.
- Minka, T., J. Winn, J. Guiver, and D. Knowles (2010). Infer.NET 2.4. Microsoft Research Cambridge. <http://research.microsoft.com/infernet>.
- Minsky, M. and S. Papert (1969). *Perceptrons*. MIT Press.
- Mitchell, T. (1997). *Machine Learning*. McGraw Hill.
- Mitchell, T. and J. Beauchamp (1988). Bayesian Variable Selection in Linear Regression. *J. of the Am. Stat. Assoc.* 83, 1023–1036.
- Mobahi, H., R. Collobert, and J. Weston (2009). Deep learning from temporal coherence in video. In *Intl. Conf. on Machine Learning*.
- Mockus, J., W. Eddy, A. Mockus, L. Mockus, and G. Reklaitis (1996). *Bayesian Heuristic Approach to Discrete and Global Optimization: Algorithms, Visualization, Software, and Applications*. Kluwer.
- Moghaddam, B., A. Gruber, Y. Weiss, and S. Avidan (2008). Sparse regression as a sparse eigenvalue problem. In *Information Theory & Applications Workshop (ITA'08)*.
- Moghaddam, B., B. Marlin, E. Khan, and K. Murphy (2009). Accelerating bayesian structural inference for non-decomposable gaussian graphical models. In *NIPS*.

- Moghaddam, B. and A. Pentland (1995). Probabilistic visual learning for object detection. In *Intl. Conf. on Computer Vision*.
- Mohamed, S., K. Heller, and Z. Ghahramani (2008). Bayesian Exponential Family PCA. In *NIPS*.
- Moler, C. (2004). *Numerical Computing with MATLAB*. SIAM.
- Morris, R. D., X. Descombes, and J. Zerubia (1996). The Ising/Potts model is not well suited to segmentation tasks. In *IEEE DSP Workshop*.
- Mosterman, P. J. and G. Biswas (1999). Diagnosis of continuous valued systems in transient operating regions. *IEEE Trans. on Systems, Man, and Cybernetics, Part A* 29(6), 554–565.
- Moulines, E., J.-F. Cardoso, and E. Gasiot (1997). Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP97)*, Munich, Germany, pp. 3617–3620.
- Muller, P., G. Parmigiani, C. Robert, and J. Rousseau (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *J. of the Am. Stat. Assoc.* 99, 990–1001.
- Mumford, D. (1994). Neuronal architectures for pattern-theoretic problems. In C. Koch and J. Davis (Eds.), *Large Scale Neuronal Theories of the Brain*. MIT Press.
- Murphy, K. (2000). Bayesian map learning in dynamic environments. In *NIPS*, Volume 12.
- Murphy, K. and M. Paskin (2001). Linear time inference in hierarchical HMMs. In *NIPS*.
- Murphy, K., Y. Weiss, and M. Jordan (1999). Loopy belief propagation for approximate inference: an empirical study. In *UAI*.
- Murphy, K. P. (1998). Filtering and smoothing in linear dynamical systems using the junction tree algorithm. Technical report, U.C. Berkeley, Dept. Comp. Sci.
- Murray, I. and Z. Ghahramani (2005). A note on the evidence and bayesian occam's razor. Technical report, Gatsby.
- Musso, C., N. Oudjane, and F. LeGland (2001). Improving regularized particle filters. In A. Doucet, J. F. G. de Freitas, and N. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*. Springer.
- Nabney, I. (2001). *NETLAB: algorithms for pattern recognition*. Springer.
- Neal, R. (1992). Connectionist learning of belief networks. *Artificial Intelligence* 56, 71–113.
- Neal, R. (1993). Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical report, Univ. Toronto.
- Neal, R. (1996). *Bayesian learning for neural networks*. Springer.
- Neal, R. (1997). Monte Carlo Implementation of Gaussian Process Models for Bayesian Regression and Classification. Technical Report 9702, U. Toronto.
- Neal, R. (1998). Erroneous Results in 'Marginal Likelihood from the Gibbs Output'. Technical report, U. Toronto.
- Neal, R. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *J. of Computational and Graphical Statistics* 9(2), 249–265.
- Neal, R. (2003a). Slice sampling. *Annals of Statistics* 31(3), 7–5–767.
- Neal, R. (2010). MCMC using Hamiltonian Dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. Chapman & Hall.
- Neal, R. and D. MacKay (1998). Likelihood-based boosting. Technical report, U. Toronto.
- Neal, R. and J. Zhang (2006). High dimensional classification Bayesian neural networks and Dirichlet diffusion trees. In I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh (Eds.), *Feature Extraction*. Springer.
- Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing* 11, 125–139.
- Neal, R. M. (2003b). Density Modeling and Clustering using Dirichlet Diffusion Trees. In J. M. Bernardo et al. (Eds.), *Bayesian Statistics* 7, pp. 619–629. Oxford University Press.
- Neal, R. M. and G. E. Hinton (1998). A new view of the EM algorithm that justifies incremental and other variants. In M. Jordan (Ed.), *Learning in Graphical Models*. MIT Press.
- Neapolitan, R. (2003). *Learning Bayesian Networks*. Prentice Hall.
- Nefian, A., L. Liang, X. Pi, X. Liu, and K. Murphy (2002). Dynamic Bayesian Networks for Audio-Visual Speech Recognition. *J. Applied Signal Processing*.
- Nemirovski, A. and D. Yudin (1978). On Cezari's convergence of the steepest descent method for approximating saddle points of convex-concave functions. *Soviet Math. Dokl.* 19.
- Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization. A basic course*. Kluwer.
- Newton, M., D. Noueiry, D. Sarkar, and P. Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5, 155–176.
- Newton, M. and A. Raftery (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *J. of Royal Stat. Soc. Series B* 56(1), 3–48.
- Ng, A., M. Jordan, and Y. Weiss (2001). On Spectral Clustering: Analysis and an algorithm. In *NIPS*.
- Ng, A. Y. and M. I. Jordan (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *NIPS-14*.
- Nickisch, H. and C. Rasmussen (2008). Approximations for binary gaussian process classification. *J. of Machine Learning Research* 9, 2035–2078.
- Nilsson, D. (1998). An efficient algorithm for finding the M most probable configurations in a probabilistic expert system. *Statistics and Computing* 8, 159–173.
- Nilsson, D. and J. Goldberger (2001). Sequentially finding the N-Best List in Hidden Markov Models. In *Intl. Joint Conf. on AI*, pp. 1280–1285.
- Nocedal, J. and S. Wright (2006). *Numerical Optimization*. Springer.

- Nowicki, K. and T. A. B. Snijders (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* 96(455), 1077–??
- Nowlan, S. and G. Hinton (1992). Simplifying neural networks by soft weight sharing. *Neural Computation* 4(4), 473–493.
- Nummiaro, K., E. Koller-Meier, and L. V. Gool (2003). An adaptive color-based particle filter. *Image and Vision Computing* 21(1), 99–110.
- Obozinski, G., B. Taskar, and M. I. Jordan (2007). Joint covariate selection for grouped classification. Technical report, UC Berkeley.
- Oh, M.-S. and J. Berger (1992). Adaptive importance sampling in Monte Carlo integration. *J. of Statistical Computation and Simulation* 41(3), 143 – 168.
- Oh, S., S. Russell, and S. Sastry (2009). Markov Chain Monte Carlo Data Association for Multi-Target Tracking. *IEEE Trans. on Automatic Control* 54(3), 481–497.
- O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *J. of Royal Stat. Soc. Series B* 40, 1–42.
- O'Hara, R. and M. Sillanpaa (2009). A Review of Bayesian Variable Selection Methods: What, How and Which. *Bayesian Analysis* 4(1), 85–118.
- Olshausen, B. A. and D. J. Field (1996). Emergence of simple cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 607–609.
- Oppel, M. (1998). A Bayesian approach to online learning. In D. Saad (Ed.), *On-line learning in neural networks*. Cambridge.
- Oppel, M. and C. Archambeau (2009). The variational Gaussian approximation revisited. *Neural Computation* 21(3), 786–792.
- Oppel, M. and D. Saad (Eds.) (2001). *Advanced mean field methods: theory and practice*. MIT Press.
- Osborne, M. R., B. Presnell, and B. A. Turlach (2000a). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* 20(3), 389–403.
- Osborne, M. R., B. Presnell, and B. A. Turlach (2000b). On the lasso and its dual. *J. Computational and graphical statistics* 9, 319–337.
- Ostendorf, M., V. Digalakis, and O. Kimball (1996). From HMMs to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Trans. on Speech and Audio Processing* 4(5), 360–378.
- Overschee, P. V. and B. D. Moor (1996). *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer Academic Publishers.
- Paatero, P. and U. Tapper (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5, 111–126.
- Padadimitriou, C. and K. Steiglitz (1982). *Combinatorial optimization: Algorithms and Complexity*. Prentice Hall.
- Paisley, J. and L. Carin (2009). Non-parametric factor analysis with beta process priors. In *Intl. Conf. on Machine Learning*.
- Palmer, S. (1999). *Vision Science: Photons to Phenomenology*. MIT Press.
- Parise, S. and M. Welling (2005). Learning in Markov Random Fields: An Empirical Study. In *Joint Statistical Meeting*.
- Park, T. and G. Casella (2008). The Bayesian Lasso. *J. of the Am. Stat. Assoc.* 103(482), 681–686.
- Parviainen, P. and M. Koivisto (2011). Ancestor relations in the presence of unobserved variables. In *Proc. European Conf. on Machine Learning*.
- Paskin, M. (2003). Thin junction tree filters for simultaneous localization and mapping. In *Intl. Joint Conf. on AI*.
- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.
- Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge Univ. Press.
- Pearl, J. and T. Verma (1991). A theory of inferred causation. In *Knowledge Representation*, pp. 441–452.
- Pe'er, D. (2005, April). Bayesian network analysis of signaling networks: a primer. *Science STKE* 281, 14.
- Peng, F., R. Jacobs, and M. Tanner (1996). Bayesian Inference in Mixtures-of-Experts and Hierarchical Mixtures-of-Experts Models With an Application to Speech Recognition. *J. of the Am. Stat. Assoc.* 91(435), 953–960.
- Petrin, G., S. Petrone, and P. Campagnoli (2009). *Dynamic linear models with R*. Springer.
- Pham, D.-T. and P. Garrat (1997). Blind separation of mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. on Signal Processing* 45(7), 1712–1725.
- Pietra, S. D., V. D. Pietra, and J. Lafferty (1997). Inducing features of random fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 19(4).
- Plackett, R. (1975). The analysis of permutations. *Applied Stat.* 24, 193–202.
- Platt, J. (1998). Using analytic QP and sparseness to speed training of support vector machines. In *NIPS*.
- Platt, J. (2000). Probabilities for sv machines. In A. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans (Eds.), *Advances in Large Margin Classifiers*. MIT Press.
- Platt, J., N. Cristianini, and J. Shawe-Taylor (2000). Large margin DAGs for multiclass classification. In *NIPS*, Volume 12, pp. 547–553.
- Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In *Proc. 3rd Intl. Workshop on Distributed Statistical Computing*.
- Polson, N. and S. Scott (2011). Data augmentation for support vector machines. *Bayesian Analysis* 6(1), 1–124.
- Pontil, M., S. Mukherjee, and F. Girosi (1998). On the Noise Model of Support Vector Machine Regression. Technical report, MIT AI Lab.
- Poon, H. and P. Domingos (2011). Sum-product networks: A new deep architecture. In *UAI*.

- Pourahmadi, M. (2004). Simultaneous Modelling of Covariance Matrices: GLM, Bayesian and Nonparametric Perspectives. Technical report, Northern Illinois University.
- Prado, R. and M. West (2010). *Time Series: Modelling, Computation and Inference*. CRC Press.
- Press, S. J. (2005). *Applied multivariate analysis, using Bayesian and frequentist methods of inference*. Dover. Second edition.
- Press, W., W. Vetterling, S. Teukolosky, and B. Flannery (1988). *Numerical Recipes in C: The Art of Scientific Computing* (Second ed.). Cambridge University Press.
- Prince, S. (2012). *Computer Vision: Models, Learning and Inference*. Cambridge.
- Pritchard, J., M. M. Stephens, and P. Donnelly (2000). Inference of population structure using multi-locus genotype data. *Genetics* 155, 945–959.
- Qi, Y. and T. Jaakkola (2008). Parameter Expanded Variational Bayesian Methods. In *NIPS*.
- Qi, Y., M. Szummer, and T. Minka (2005). Bayesian Conditional Random Fields. In *10th Intl. Workshop on AI/Statistics*.
- Quinlan, J. (1990). Learning logical definitions from relations. *Machine Learning* 5, 239–266.
- Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning* 1, 81–106.
- Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning*. Morgan Kaufman.
- Quinonero-Candela, J., C. Rasmussen, and C. Williams (2007). Approximation methods for gaussian process regression. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston (Eds.), *Large Scale Kernel Machines*, pp. 203–223. MIT Press.
- Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. of the IEEE* 77(2), 257–286.
- Rai, P. and H. Daume (2009). Multi-label prediction via sparse infinite CCA. In *NIPS*.
- Raiffa, H. (1968). *Decision Analysis*. Addison Wesley.
- Raina, R., A. Madhavan, and A. Ng (2009). Large-scale deep unsupervised learning using graphics processors. In *Intl. Conf. on Machine Learning*.
- Raina, R., A. Ng, and D. Koller (2005). Transfer learning by constructing informative priors. In *NIPS*.
- Rajaraman, A. and J. Ullman (2010). *Mining of massive datasets*. Self-published.
- Rajaraman, A. and J. Ullman (2011). *Mining of massive datasets*. Cambridge.
- Rakotomamonjy, A., F. Bach, S. Canu, and Y. Grandvalet (2008). SimpleMKL. *J. of Machine Learning Research* 9, 2491–2521.
- Ramage, D., D. Hall, R. Nallapati, and C. Manning (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP*.
- Ramage, D., C. Manning, and S. Dumais (2011). Partially Labeled Topic Models for Interpretable Text Mining. In *Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining*.
- Ramaswamy, S., P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander, and T. Golub (2001). Multiclass cancer diagnosis using tumor gene expression signature. *Proc. of the National Academy of Science, USA* 98, 15149–15154.
- Ranzato, M. and G. Hinton (2010). Modeling pixel means and covariances using factored third-order Boltzmann machines. In *CVPR*.
- Ranzato, M., F.-J. Huang, Y.-L. Boureau, and Y. LeCun (2007). Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition. In *CVPR*.
- Ranzato, M., C. Poultney, S. Chopra, and Y. LeCun (2006). Efficient learning of sparse representations with an energy-based model. In *NIPS*.
- Rao, A. and K. Rose (2001, February). Deterministically Annealed Design of Hidden Markov Model Speech Recognizers. *IEEE Trans. on Speech and Audio Proc.* 9(2), 111–126.
- Rasmussen, C. (2000). The infinite gaussian mixture model. In *NIPS*.
- Rasmussen, C. E. and J. Quiñonero-Candela (2005). Healing the relevance vector machine by augmentation. In *Intl. Conf. on Machine Learning*, pp. 689–696.
- Rasmussen, C. E. and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*. MIT Press.
- Ratsch, G., T. Onoda, and K. Muller (2001). Soft margins for adaboost. *Machine Learning* 42, 287–320.
- Ratnay, M., O. Stegle, K. Sharp, and J. Winn (2009). Inference algorithms and learning theory for Bayesian sparse factor analysis. In *Proc. Intl. Workshop on Statistical Mechanical Informatics*.
- Rauch, H. E., F. Tung, and C. T. Striebel (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal* 3(8), 1445–1450.
- Ravikumar, P., J. Lafferty, H. Liu, and L. Wasserman (2009). Sparse Additive Models. *J. of Royal Stat. Soc. Series B* 71(5), 1009–1030.
- Raydan, M. (1997). The barzilai and borwein gradient method for the large scale unconstrained minimization problem. *SIAM J. on Optimization* 7(1), 26–33.
- Rennie, J. (2004). Why sums are bad. Technical report, MIT.
- Rennie, J., L. Shih, J. Teevan, and D. Karger (2003). Tackling the poor assumptions of naive Bayes text classifiers. In *Intl. Conf. on Machine Learning*.
- Reshed, D., Y. Reshef, H. Finucane, S. Grossman, G. McVean, P. Turnbaugh, E. Lander, M. Mitzenmacher, and P. Sabeti (2011, December). Detecting novel associations in large data sets. *Science* 334, 1518–1524.
- Resnick, S. I. (1992). *Adventures in Stochastic Processes*. Birkhauser.
- Rice, J. (1995). *Mathematical statistics and data analysis*. Duxbury. 2nd edition.

- Richardson, S. and P. Green (1997). On Bayesian Analysis of Mixtures With an Unknown Number of Components. *J. of Royal Stat. Soc. Series B* 59, 731–758.
- Riesenhuber, M. and T. Poggio (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience* 2, 1019–1025.
- Rish, I., G. Grabarnik, G. Cecchi, F. Pereira, and G. Gordon (2008). Closed-form supervised dimensionality reduction with generalized linear models. In *Intl. Conf. on Machine Learning*.
- Ristic, B., S. Arulampalam, and N. Gordon (2004). *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House Radar Library.
- Robert, C. (1995). Simulation of truncated normal distributions. *Statistics and computing* 5, 121–125.
- Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods*. Springer. 2nd edition.
- Roberts, G. and J. Rosenthal (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* 16, 351–367.
- Roberts, G. O. and S. K. Sahu (1997). Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *J. of Royal Stat. Soc. Series B* 59(2), 291–317.
- Robinson, R. W. (1973). Counting labeled acyclic digraphs. In F. Harary (Ed.), *New Directions in the Theory of Graphs*, pp. 239–273. Academic Press.
- Roch, S. (2006). A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Trans. Comp. Bio. Bioinformatics* 3(1).
- Rodriguez, A. and K. Ghosh (2011). Modeling relational data through nested partition models. *Biometrika*. To appear.
- Rose, K. (1998, November). Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE* 80, 2210–2239.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review* 65(6), 386–408.
- Ross, S. (1989). *Introduction to Probability Models*. Academic Press.
- Rosset, S., J. Zhu, and T. Hastie (2004). Boosting as a regularized path to a maximum margin classifier. *J. of Machine Learning Research* 5, 941–973.
- Rossi, P., G. Allenby, and R. McCulloch (2006). *Bayesian Statistics and Marketing*. Wiley.
- Roth, D. (1996, Apr). On the hardness of approximate reasoning. *Artificial Intelligence* 82(1-2), 273–302.
- Rother, C., P. Kohli, W. Feng, and J. Jia (2009). Minimizing sparse higher order energy functions of discrete variables. In *CVPR*, pp. 1382–1389.
- Rouder, J., P. Speckman, D. Sun, and R. Morey (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* 16(2), 225–237.
- Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statistics* 29, 391–411.
- Roweis, S. (1997). EM algorithms for PCA and SPCA. In *NIPS*.
- Rubin, D. (1998). Using the SIR algorithm to simulate posterior distributions. In *Bayesian Statistics* 3.
- Rue, H. and L. Held (2005). *Gaussian Markov Random Fields: Theory and Applications*, Volume 104 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.
- Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations. *J. of Royal Stat. Soc. Series B* 71, 319–392.
- Rumelhart, D., G. Hinton, and R. Williams (1986). Learning internal representations by error propagation. In D. Rumelhart, J. McClelland, and the PDD Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. MIT Press.
- Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge University Press.
- Rush, A. M. and M. Collins (2012). A tutorial on Lagrangian relaxation and dual decomposition for NLP. Technical report, Columbia U.
- Russell, S., J. Binder, D. Koller, and K. Kanazawa (1995). Local learning in probabilistic networks with hidden variables. In *Intl. Joint Conf. on AI*.
- Russell, S. and P. Norvig (1995). *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.
- Russell, S. and P. Norvig (2002). *Artificial Intelligence: A Modern Approach*. Prentice Hall. 2nd edition.
- Russell, S. and P. Norvig (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall. 3rd edition.
- S. and M. Black (2009, April). Fields of experts. *Intl. J. Computer Vision* 82(2), 205–229.
- Sachs, K., O. Perez, D. Pe’er, D. Lauffenburger, and G. Nolan (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308.
- Sahami, M. and T. Heilman (2006). A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In *WWW conferenec*.
- Salakhutdinov, R. (2009). *Deep Generative Models*. Ph.D. thesis, U. Toronto.
- Salakhutdinov, R. and G. Hinton (2009). Deep Boltzmann machines. In *AI/Statistics*, Volume 5, pp. 448–455.
- Salakhutdinov, R. and G. Hinton (2010). Replicated Softmax: an Undirected Topic Model. In *NIPS*.
- Salakhutdinov, R. and H. Larochelle (2010). Efficient Learning of Deep Boltzmann Machines. In *AI/Statistics*.
- Salakhutdinov, R. and A. Mnih (2008). Probabilistic matrix factorization. In *NIPS*, Volume 20.

- Salakhutdinov, R. and S. Roweis (2003). Adaptive overrelaxed bound optimization methods. In *Proceedings of the International Conference on Machine Learning*, Volume 20, pp. 664–671.
- Salakhutdinov, R., J. Tenenbaum, and A. Torralba (2011). Learning To Learn with Compound HD Models. In *NIPS*.
- Salakhutdinov, R. R., A. Mnih, and G. E. Hinton (2007). Restricted boltzmann machines for collaborative filtering. In *Intl. Conf. on Machine Learning*, Volume 24, pp. 791–798.
- Salojärvi, J., K. Puolamäki, and S. Klaski (2005). On discriminative joint density modeling. In *Proc. European Conf. on Machine Learning*.
- Sampson, F. (1968). *A Novitiate in a Period of Change: An Experimental and Case Study of Social Relationships*. Ph.D. thesis, Cornell.
- Santner, T., B. Williams, and W. Notz (2003). *The Design and Analysis of Computer Experiments*. Springer.
- Sarkar, J. (1991). One-armed bandit problems with covariates. *The Annals of Statistics* 19(4), 1978–2002.
- Sato, M. and S. Ishii (2000). On-line EM algorithm for the normalized Gaussian network. *Neural Computation* 12, 407–432.
- Saul, L., T. Jaakkola, and M. Jordan (1996). Mean Field Theory for Sigmoid Belief Networks. *J. of AI Research* 4, 61–76.
- Saul, L. and M. Jordan (1995). Exploiting tractable substructures in intractable networks. In *NIPS*, Volume 8.
- Saul, L. and M. Jordan (2000). Attractor dynamics in feedforward neural networks. *Neural Computation* 12, 1313–1335.
- Saunders, C., J. Shawe-Taylor, and A. Vinokourov (2003). String Kernels, Fisher Kernels and Finite State Automata. In *NIPS*.
- Savage, R., K. Heller, Y. Xi, Z. Ghahramani, W. Truman, M. Grant, K. Denby, and D. Wild (2009). R/BHC: fast Bayesian hierarchical clustering for microarray data. *BMC Bioinformatics* 10(242).
- Schaefer, J. and K. Strimmer (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol* 4(32).
- Schapire, R. (1990). The strength of weak learnability. *Machine Learning* 5, 197–227.
- Schapire, R. and Y. Freund (2012). *Boosting: Foundations and Algorithms*. MIT Press.
- Schapire, R., Y. Freund, P. Bartlett, and W. Lee (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. *Annals of Statistics* 5, 1651–1686.
- Scharstein, D. and R. Szeliski (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Intl. J. Computer Vision* 47(1), 7–42.
- Schaul, T., S. Zhang, and Y. LeCun (2012). No more pesky learning rates. Technical report, Courant Institute of Mathematical Sciences.
- Schmee, J. and G. Hahn (1979). A simple method for regression analysis with censored data. *Technometrics* 21, 417–432.
- Schmidt, M. (2010). *Graphical model structure learning with L1 regularization*. Ph.D. thesis, UBC.
- Schmidt, M., G. Fung, and R. Rosales (2009). Optimization methods for $\ell - 1$ regularization. Technical report, U. British Columbia.
- Schmidt, M. and K. Murphy (2009). Modeling Discrete Interventional Data using Directed Cyclic Graphical Models. In *UAI*.
- Schmidt, M., K. Murphy, G. Fung, and R. Rosales (2008). Structure Learning in Random Fields for Heart Motion Abnormality Detection. In *CVPR*.
- Schmidt, M., A. Niculescu-Mizil, and K. Murphy (2007). Learning Graphical Model Structure using L1-Regularization Paths. In *AAAI*.
- Schmidt, M., E. van den Berg, M. Friedlander, and K. Murphy (2009). Optimizing Costly Functions with Simple Constraints: A Limited-Memory Projected Quasi-Newton Algorithm. In *AI & Statistics*.
- Schniter, P., L. C. Potter, and J. Ziniel (2008). Fast Bayesian Matching Pursuit: Model Uncertainty and Parameter Estimation for Sparse Linear Models. Technical report, U. Ohio. Submitted to IEEE Trans. on Signal Processing.
- Schnitzspan, P., S. Roth, and B. Schiele (2010). Automatic discovery of meaningful object parts with latent CRFs. In *CVPR*.
- Schoelkopf, B. and A. Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.
- Schoelkopf, B., A. Smola, and K.-R. Müller (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* 10, 1299 – 1319.
- Schraudolph, N. N., J. Yu, and S. Günter (2007). A Stochastic Quasi-Newton Method for Online Convex Optimization. In *AI/Statistics*, pp. 436–443.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6(2), 461–464.
- Schwarz, R. and Y. Chow (1990). The n-best algorithm: an efficient and exact procedure for finding the n most likely hypotheses. In *Intl. Conf. on Acoustics, Speech and Signal Proc.*
- Schweikerta, G., A. Zien, G. Zeller, J. Behr, C. Dieterich, C. Ong, P. Philips, F. D. Bona, L. Hartmann, A. Böhlen, N. KrÄjger, S. Sonnenburg, and G. RÄdtsch (2009). mGene: Accurate SVM-based Gene Finding with an Application to Nematode Genomes. *Genome Research*, 19, 2133–2143.
- Scott, D. (1979). On optimal and data-based histograms. *Biometrika* 66(3), 605–610.
- Scott, J. G. and C. M. Carvalho (2008). Feature-inclusion stochastic search for gaussian graphical models. *J. of Computational and Graphical Statistics* 17(4), 790–808.
- Scott, S. (2009). Data augmentation, frequentist estimation, and the bayesian analysis of multinomial logit models. *Statistical Papers*.
- Scott, S. (2010). A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry* 26, 639–658.

- Sedgewick, R. and K. Wayne (2011). *Algorithms*. Addison Wesley.
- Seeger, M. (2008). Bayesian Inference and Optimal Design in the Sparse Linear Model. *J. of Machine Learning Research* 9, 759–813.
- Seeger, M. and H. Nickish (2008). Compressed sensing and bayesian experimental design. In *Intl. Conf. on Machine Learning*.
- Segal, D. (2011, 12 February). The dirty little secrets of search. *New York Times*.
- Seide, F., G. Li, and D. Yu (2011). Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. In *Interspeech*.
- Sejnowski, T. and C. Rosenberg (1987). Parallel networks that learn to pronounce english text. *Complex Systems* 1, 145–168.
- Sellke, T., M. J. Bayarri, and J. Berger (2001). Calibration of p Values for Testing Precise Null Hypotheses. *The American Statistician* 55(1), 62–71.
- Serre, T., L. Wolf, and T. Poggio (2005). Object recognition with features inspired by visual cortex. In *CVPR*, pp. 994–1000.
- Shachter, R. (1998). Bayes-ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *UAI*.
- Shachter, R. and C. R. Kenley (1989). Gaussian influence diagrams. *Managment Science* 35(5), 527–550.
- Shachter, R. D. and M. A. Peot (1989). Simulation approaches to general probabilistic inference on belief networks. In *UAI*, Volume 5.
- Shafer, G. R. and P. P. Shenoy (1990). Probability propagation. *Annals of Mathematics and AI* 2, 327–352.
- Shafto, P., C. Kemp, V. Mansinghka, M. Gordon, and J. B. Tenenbaum (2006). Learning cross-cutting systems of categories. In *Cognitive Science Conference*.
- Shahaf, D., A. Chechetka, and C. Guestrin (2009). Learning Thin Junction Trees via Graph Cuts. In *AISTATS*.
- Shalev-Shwartz, S., Y. Singer, and N. Srebro (2007). Pegasos: primal estimated sub-gradient solver for svm. In *Intl. Conf. on Machine Learning*.
- Shalizi, C. (2009). Cs 36-350 lecture 10: Principal components: mathematics, example, interpretation.
- Shan, H. and A. Banerjee (2010). Residual Bayesian co-clustering for matrix approximation. In *SIAM Intl. Conf. on Data Mining*.
- Shawe-Taylor, J. and N. Cristianini (2004). *Kernel Methods for Pattern Analysis*. Cambridge.
- Sheng, Q., Y. Moreau, and B. D. Moor (2003). Biclustering Microarray data by Gibbs sampling. *Bioinformatics* 19, ii196–ii205.
- Shi, J. and J. Malik (2000). Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.
- Shoham, Y. and K. Leyton-Brown (2009). *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.
- Shotton, J., A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake (2011). Real-time human pose recognition in parts from a single depth image. In *CVPR*.
- Shwe, M., B. Middleton, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper (1991). Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base. *Methods. Inf. Med* 30(4), 241–255.
- Siddiqi, S., B. Boots, and G. Gordon (2007). A constraint generation approach to learning stable linear dynamical systems. In *NIPS*.
- Siepel, A. and D. Haussler (2003). Combining phylogenetic and hidden markov models in biosequence analysis. In *Proc. 7th Intl. Conf. on Computational Molecular Biology (RECOMB)*.
- Silander, T., P. Kontkanen, and P. Myllymäki (2007). On Sensitivity of the MAP Bayesian Network Structure to the Equivalent Sample Size Parameter. In *UAI*, pp. 360–367.
- Silander, T. and P. Myllymäki (2006). A simple approach for finding the globally optimal Bayesian network structure. In *UAI*.
- Sill, J., G. Takacs, L. Mackey, and D. Lin (2009). Feature-weighted linear stacking. Technical report, .
- Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Annals of Statistics* 12(3), 898–916.
- Simard, P., D. Steinkraus, and J. Platt (2003). Best practices for convolutional neural networks applied to visual document analysis. In *Intl. Conf. on Document Analysis and Recognition (ICDAR)*.
- Simon, D. (2006). *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. Wiley.
- Singliar, T. and M. Hauskrecht (2006). Noisy-OR Component Analysis and its Application to Link Analysis. *J. of Machine Learning Research* 7.
- Smidl, V. and A. Quinn (2005). *The Variational Bayes Method in Signal Processing*. Springer.
- Smith, A. F. M. and A. E. Gelfand (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician* 46(2), 84–88.
- Smith, R. and P. Cheeseman (1986). On the representation and estimation of spatial uncertainty. *Intl. J. Robotics Research* 5(4), 56–68.
- Smith, V., J. Yu, T. Smulders, A. Hartemink, and E. Jarvis (2006). Computational Inference of Neural Information Flow Networks. *PLOS Computational Biology* 2, 1436–1439.
- Smolensky, P. (1986). Information processing in dynamical systems: foundations of harmony theory. In D. Rumelhart and J. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume I*. McGraw-Hill.
- Smyth, P., D. Heckerman, and M. I. Jordan (1997). Probabilistic independence networks for hidden Markov probability models. *Neural Computation* 9(2), 227–269.
- Sohl-Dickstein, J., P. Battaglino, and M. DeWeese (2011). In *Intl. Conf. on Machine Learning*.

- Sollich, P. (2002). Bayesian methods for support vector machines: evidence and predictive class probabilities. *Machine Learning* 46, 21–52.
- Sontag, D., A. Globerson, and T. Jaakkola (2011). Introduction to dual decomposition for inference. In S. Sra, S. Nowozin, and S. J. Wright (Eds.), *Optimization for Machine Learning*. MIT Press.
- Sorenson, H. and D. Alspach (1971). Recursive Bayesian estimation using Gaussian sums. *Automatica* 7, 465–479.
- Soussen, C., J. Iier, D. Brie, and J. Duan (2010). From Bernoulli-Gaussian deconvolution to sparse signal restoration. Technical report, Centre de Recherche en Automatique de Nancy.
- Spaan, M. and N. Vlassis (2005). Perseus: Randomized Point-based Value Iteration for POMDPs. *J. of AI Research* 24, 195–220.
- Spall, J. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley.
- Speed, T. (2011, December). A correlation for the 21st century. *Science* 334, 152–1503.
- Speed, T. and H. Kiiveri (1986). Gaussian Markov distributions over finite graphs. *Annals of Statistics* 14(1), 138–150.
- Spiegelhalter, D. J. and S. L. Lauritzen (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* 20.
- Spirites, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search*. MIT Press. 2nd edition.
- Srebro, N. (2001). Maximum Likelihood Bounded Tree-Width Markov Networks. In *UAI*.
- Srebro, N. and T. Jaakkola (2003). Weighted low-rank approximations. In *Intl. Conf. on Machine Learning*.
- Steinbach, M., G. Karypis, and V. Kumar (2000). A comparison of document clustering techniques. In *KDD Workshop on Text Mining*.
- Stephens, M. (2000). Dealing with label-switching in mixture models. *J. Royal Statistical Society, Series B* 62, 795–809.
- Stern, D., R. Herbrich, and T. Graepel (2009). Matchbox: Large Scale Bayesian Recommendations. In *Proc. 18th. Intl. World Wide Web Conference*.
- Steyvers, M. and T. Griffiths (2007). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch (Eds.), *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.
- Stigler, S. (1986). *The history of statistics*. Harvard University press.
- Stolcke, A. and S. M. Omohundro (1992). Hidden Markov Model Induction by Bayesian Model Merging. In *NIPS-5*.
- Stoyanov, V., A. Ropson, and J. Eisner (2011). Empirical risk minimization of graphical model parameters given approximate inference, decoding, and model structure. In *AI/Statistics*.
- Sudderth, E. (2006). *Graphical Models for Visual Object Recognition and Tracking*. Ph.D. thesis, MIT.
- Sudderth, E. and W. Freeman (2008, March). Signal and Image Processing with Belief Propagation. *IEEE Signal Processing Magazine*.
- Sudderth, E., A. Ihler, W. Freeman, and A. Willsky (2003). Nonparametric Belief Propagation. In *CVPR*.
- Sudderth, E., A. Ihler, M. Isard, W. Freeman, and A. Willsky (2010). Nonparametric Belief Propagation. *Comm. of the ACM* 53(10).
- Sudderth, E. and M. Jordan (2008). Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes. In *NIPS*.
- Sudderth, E., M. Wainwright, and A. Willsky (2008). Loop series and bethe variational bounds for attractive graphical models. In *NIPS*.
- Sun, J., N. Zheng, and H. Shum (2003). Stereo matching using belief propagation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 25(7), 787–800.
- Sun, L., S. Ji, S. Yu, and J. Ye (2009). On the equivalence between canonical correlation analysis and orthonormalized partial least squares. In *Intl. Joint Conf. on AI*.
- Sunehag, P., J. Trunpf, S. V. N. Vishwanathan, and N. N. Schraudolph (2009). Variable Metric Stochastic Approximation Theory. In *AI/Statistics*, pp. 560–566.
- Sutton, C. and A. McCallum (2007). Improved Dynamic Schedules for Belief Propagation. In *UAI*.
- Sutton, R. and A. Barto (1998). *Reinforcement Learning: An Introduction*. MIT Press.
- Swendsen, R. and J.-S. Wang (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters* 58, 86–88.
- Swersky, K., B. Chen, B. Marlin, and N. de Freitas (2010). A Tutorial on Stochastic Approximation Algorithms for Training Restricted Boltzmann Machines and Deep Belief Nets. In *Information Theory and Applications (ITA) Workshop*.
- Szeliski, R. (2010). *Computer Vision: Algorithms and Applications*. Springer.
- Szeliski, R., R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother (2008). A Comparative Study of Energy Minimization Methods for Markov Random Fields with Smoothness-Based Priors. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 30(6), 1068–1080.
- Szepesvari, C. (2010). *Algorithms for Reinforcement Learning*. Morgan Claypool.
- Taleb, N. (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House.
- Talhouk, A., K. Murphy, and A. Doucet (2011). Efficient Bayesian Inference for Multivariate Probit Models with Sparse Inverse Correlation Matrices. *J. Comp. Graph. Statist.*
- Tanner, M. (1996). *Tools for statistical inference*. Springer.
- Tanner, M. and W. Wong (1987). The calculation of posterior distributions by data augmentation. *J. of the Am. Stat. Assoc.* 82(398), 528–540.

- Tarlow, D., I. Givoni, and R. Zemel (2010). Hop-map: efficient message passing with high order potentials. In *AI/Statistics*.
- Taskar, B., C. Guestrin, and D. Koller (2003). Max-margin markov networks. In *NIPS*.
- Taskar, B., D. Klein, M. Collins, D. Koller, and C. Manning (2004). Max-margin parsing. In *Proc. Empirical Methods in Natural Language Processing*.
- Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proc. of the Assoc. for Computational Linguistics*, pp. 985–992.
- Teh, Y.-W., M. Jordan, M. Beal, and D. Blei (2006). Hierarchical Dirichlet processes. *J. of the Am. Stat. Assoc.* 101(476), 1566–1581.
- Tenenbaum, J. (1999). *A Bayesian framework for concept learning*. Ph.D. thesis, MIT.
- Tenenbaum, J. B. and F. Xu (2000). Word learning as bayesian inference. In *Proc. 22nd Annual Conf. of the Cognitive Science Society*.
- Theocharous, G., K. Murphy, and L. Kaelbling (2004). Representing hierarchical POMDPs as DBNs for multi-scale robot localization. In *IEEE Intl. Conf. on Robotics and Automation*.
- Thiesson, B., C. Meek, D. Chickering, and D. Heckerman (1998). Learning mixtures of DAG models. In *UAI*.
- Thomas, A. and P. Green (2009). Enumerating the decomposable neighbours of a decomposable graph under a simple perturbation scheme. *Comp. Statistics and Data Analysis* 53, 1232–1238.
- Thrun, S., W. Burgard, and D. Fox (2006). *Probabilistic Robotics*. MIT Press.
- Thrun, S., M. Montemerlo, D. Koller, B. Wegbreit, J. Nieto, and E. Nebot (2004). Fastslam: An efficient solution to the simultaneous localization and mapping problem with unknown data association. *J. of Machine Learning Research* 2004.
- Thrun, S. and L. Pratt (Eds.) (1997). *Learning to learn*. Kluwer.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B* 58(1), 267–288.
- Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a dataset via the gap statistic. *J. of Royal Stat. Soc. Series B* 32(2), 411–423.
- Tieleman, T. (2008). Training restricted Boltzmann machines using approximations to the likelihood gradient. In *Proceedings of the 25th international conference on Machine learning*, pp. 1064–1071. ACM New York, NY, USA.
- Ting, J., A. D'Souza, S. Vijayakumar, and S. Schaal (2010). Efficient learning and feature selection in high-dimensional regression. *Neural Computation* 22(4), 831–886.
- Tipping, M. (1998). Probabilistic visualization of high-dimensional binary data. In *NIPS*.
- Tipping, M. (2001). Sparse bayesian learning and the relevance vector machine. *J. of Machine Learning Research* 1, 211–244.
- Tipping, M. and C. Bishop (1999). Probabilistic principal component analysis. *J. of Royal Stat. Soc. Series B* 2(3), 611–622.
- Tipping, M. and A. Faul (2003). Fast marginal likelihood maximisation for sparse bayesian models. In *AI/Statistics*.
- Tishby, N., F. Pereira, and W. Biale (1999). The information bottleneck method. In *The 37th annual Allerton Conf. on Communication, Control, and Computing*, pp. 368–377.
- Tomas, M., D. Anoop, K. Stefan, B. Lukas, and C. Jan (2011). Empirical evaluation and combination of advanced language modeling techniques. In *Proc. 12th Annual Conf. of the Intl. Speech Communication Association (INTERSPEECH)*.
- Torralba, A., R. Fergus, and Y. Weiss (2008). Small codes and large image databases for recognition. In *CVPR*.
- Train, K. (2009). *Discrete choice methods with simulation*. Cambridge University Press. Second edition.
- Tseng, P. (2008). On Accelerated Proximal Gradient Methods for Convex-Concave Optimization. Technical report, U. Washington.
- Tsochantaridis, I., T. Joachims, T. Hofmann, and Y. Altun (2005, September). Large margin methods for structured and interdependent output variables. *J. of Machine Learning Research* 6, 1453–1484.
- Tu, Z. and S. Zhu (2002). Image Segmentation by Data-Driven Markov Chain Monte Carlo. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24(5), 657–673.
- Turian, J., L. Ratinov, and Y. Bengio (2010). Word representations: a simple and general method for semi-supervised learning. In *Proc. ACL*.
- Turlach, B., W. Venables, and S. Wright (2005). Simultaneous variable selection. *Technometrics* 47(3), 349–363.
- Turner, R., P. Berkes, M. Sahani, and D. Mackay (2008). Counterexamples to variational free energy compactness folk theorems. Technical report, U. Cambridge.
- Ueda, N. and R. Nakano (1998). Deterministic annealing EM algorithm. *Neural Networks* 11, 271–282.
- Usunier, N., D. Buffoni, and P. Gallinari (2009). Ranking with ordered weighted pairwise classification.
- Vaithyanathan, S. and B. Dom (1999). Model selection in unsupervised learning with applications to document clustering. In *Intl. Conf. on Machine Learning*.
- van der Merwe, R., A. Doucet, N. de Freitas, and E. Wan (2000). The unscented particle filter. In *NIPS-13*.
- van Dyk, D. and X.-L. Meng (2001). The Art of Data Augmentation. *J. Computational and Graphical Statistics* 10(1), 1–50.
- Vandenberghe, L. (2006). Applied numerical computing: Lecture notes.
- Vandenberghe, L. (2011). Ee236c - optimization methods for large-scale systems.
- Vanhatalo, J. (2010). *Speeding up the inference in Gaussian process models*. Ph.D. thesis, Helsinki Univ. Technology.

- Vanhatalo, J., V. Pietiläinen, and A. Vehtari (2010). Approximate inference for disease mapping with sparse gaussian processes. *Statistics in Medicine* 29(15), 1580–1607.
- Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.
- Vapnik, V., S. Golowich, and A. Smola (1997). Support vector method for function approximation, regression estimation, and signal processing. In *NIPS*.
- Varian, H. (2011). Structural time series in R: a Tutorial. Technical report, Google.
- Verma, T. and J. Pearl (1990). Equivalence and synthesis of causal models. In *UAI*.
- Viinikanoja, J., A. Klami, and S. Kaski (2010). Variational Bayesian Mixture of Robust CCA Models. In *Proc. European Conf. on Machine Learning*.
- Vincent, P. (2011). A Connection between Score Matching and Denoising Autoencoders. *Neural Computation* 23(7), 1661–1674.
- Vincent, P., H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol (2010). Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. of Machine Learning Research* 11, 3371–3408.
- Vinh, N., J. Epps, and J. Bailey (2009). Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Intl. Conf. on Machine Learning*.
- Vinyals, M., J. Cerquides, J. Rodriguez-Aguilar, and A. Farinelli (2010). Worst-case bounds on the quality of max-product fixed-points. In *NIPS*.
- Viola, P. and M. Jones (2001). Rapid object detection using a boosted cascade of simple classifiers. In *CVPR*.
- Virtanen, S. (2010). Bayesian exponential family projections. Master's thesis, Aalto University.
- Vishwanathan, S. V. N. and A. Smola (2003). Fast kernels for string and tree matching. In *NIPS*.
- Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. on Information Theory* 13(2), 260–269.
- von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing* 17(4), 395–416.
- Wagenmakers, E.-J., R. Wetzels, D. Borsboom, and H. van der Maas (2011). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi. *Journal of Personality and Social Psychology*.
- Wagner, D. and F. Wagner (1993). Between min cut and graph bisection. In *Proc. 18th Intl. Symp. on Math. Found. of Comp. Sci.*, pp. 744–750.
- Wainwright, M., T. Jaakkola, and A. Willsky (2001). Tree-based reparameterization for approximate estimation on loopy graphs. In *NIPS-14*.
- Wainwright, M., T. Jaakkola, and A. Willsky (2005). A new class of upper bounds on the log partition function. *IEEE Trans. Info. Theory* 51(7), 2313–2335.
- Wainwright, M., P. Ravikumar, and J. Lafferty (2006). Inferring graphical model structure using ℓ_1 -regularized pseudo-likelihood. In *NIPS*.
- Wainwright, M. J., T. S. Jaakkola, and A. S. Willsky (2003). Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Trans. on Information Theory* 49(5), 1120–1146.
- Wainwright, M. J. and M. I. Jordan (2008a). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* 1–2, 1–305.
- Wainwright, M. J. and M. I. Jordan (2008b). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning* 1–2, 1–305.
- Wallach, H., I. Murray, R. Salakhutdinov, and D. Mimno (2009). Evaluation methods for topic models. In *Intl. Conf. on Machine Learning*.
- Wan, E. A. and R. V. der Merwe (2001). The Unscented Kalman Filter. In S. Haykin (Ed.), *Kalman Filtering and Neural Networks*. Wiley.
- Wand, M. (2009). Semiparametric regression and graphical models. *Aust. N. Z. J. Stat.* 51(1), 9–41.
- Wand, M. P., J. T. Ormerod, S. A. Padoan, and R. Fruhwirth (2011). Mean Field Variational Bayes for Elaborate Distributions. *Bayesian Analysis* 6(4), 847 – 900.
- Wang, C. (2007). Variational Bayesian Approach to Canonical Correlation Analysis. *IEEE Trans. on Neural Networks* 18(3), 905–910.
- Wasserman, L. (2004). *All of statistics. A concise course in statistical inference*. Springer.
- Wei, G. and M. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. of the Am. Stat. Assoc.* 85(411), 699–704.
- Weinberger, K., A. Dasgupta, J. Attenberg, J. Langford, and A. Smola (2009). Feature hashing for large scale multitask learning. In *Intl. Conf. on Machine Learning*.
- Weiss, D., B. Sapp, and B. Taskar (2010). Sidestepping intractable inference with structured ensemble cascades. In *NIPS*.
- Weiss, Y. (2000). Correctness of local probability propagation in graphical models with loops. *Neural Computation* 12, 1–41.
- Weiss, Y. (2001). Comparing the mean field method and belief propagation for approximate inference in MRFs. In Saad and Oppor (Eds.), *Advanced Mean Field Methods*. MIT Press.
- Weiss, Y. and W. T. Freeman (1999). Correctness of belief propagation in Gaussian graphical models of arbitrary topology. In *NIPS-12*.
- Weiss, Y. and W. T. Freeman (2001a). Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation* 13(10), 2173–2200.
- Weiss, Y. and W. T. Freeman (2001b). On the optimality of solutions of the max-product belief propagation algorithm in arbitrary graphs. *IEEE Trans. Information Theory, Special Issue on Codes on Graphs and Iterative Algorithms* 47(2), 723–735.
- Weiss, Y., A. Torralba, and R. Fergus (2008). Spectral hashing. In *NIPS*.

- Welling, M., C. Chemudugunta, and N. Sutter (2008). Deterministic latent variable models and their pitfalls. In *Intl. Conf. on Data Mining*.
- Welling, M., T. Minka, and Y. W. Teh (2005). Structured region graphs: Morphing EP into GBP. In *UAI*.
- Welling, M., M. Rosen-Zvi, and G. Hinton (2004). Exponential family harmoniums with an application to information retrieval. In *NIPS-14*.
- Welling, M. and C. Sutton (2005). Learning in Markov random fields with contrastive free energies. In *Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS)*.
- Welling, M. and Y.-W. Teh (2001). Belief optimization for binary networks: a stable alternative to loopy belief propagation. In *UAI*.
- Werbos, P. (1974). *Beyond regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. Ph.D. thesis, Harvard.
- West, M. (1987). On scale mixtures of normal distributions. *Biometrika* 74, 646–648.
- West, M. (2003). Bayesian Factor Regression Models in the "Large p, Small n" Paradigm. *Bayesian Statistics 7*.
- West, M. and J. Harrison (1997). *Bayesian forecasting and dynamic models*. Springer.
- Weston, J., S. Bengio, and N. Usunier (2010). Large Scale Image Annotation: Learning to Rank with Joint Word-Image Embeddings. In *Proc. European Conf. on Machine Learning*.
- Weston, J., F. Ratle, and R. Collobert (2008). Deep Learning via Semi-Supervised Embedding. In *Intl. Conf. on Machine Learning*.
- Weston, J. and C. Watkins (1999). Multi-class support vector machines. In *ESANN*.
- Wiering, M. and M. van Otterlo (Eds.) (2012). *Reinforcement learning: State-of-the-art*. Springer.
- Wilkinson, D. and S. Yeung (2002). Conditional simulation from highly structured gaussian systems with application to blocking-mcmc for the bayesian analysis of very large linear models. *Statistics and Computing* 12, 287–300.
- Williams, C. (1998). Computation with infinite networks. *Neural Computation* 10(5), 1203–1216.
- Williams, C. (2000). A MCMC approach to Hierarchical Mixture Modelling. In S. A. Solla, T. K. Leen, and K.-R. Müller (Eds.), *NIPS*. MIT Press.
- Williams, C. (2002). On a Connection between Kernel PCA and Metric Multidimensional Scaling. *Machine Learning J.* 46(1).
- Williams, O. and A. Fitzgibbon (2006). Gaussian process implicit surfaces. In *Gaussian processes in practice*.
- Williamson, S. and Z. Ghahramani (2008). Probabilistic models for data combination in recommender systems. In *NIPS Workshop on Learning from Multiple Sources*.
- Winn, J. and C. Bishop (2005). Variational message passing. *J. of Machine Learning Research* 6, 661–694.
- Wipf, D. and S. Nagarajan (2007). A new view of automatic relevancy determination. In *NIPS*.
- Wipf, D. and S. Nagarajan (2010, April). Iterative Reweighted $\ell-1$ and $\ell-2$ Methods for Finding Sparse Solutions. *J. of Selected Topics in Signal Processing (Special Issue on Compressive Sensing)* 4(2).
- Wipf, D., B. Rao, and S. Nagarajan (2010). Latent variable bayesian models for promoting sparsity. *IEEE Transactions on Information Theory*.
- Witten, D., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* 10(3), 515–534.
- Wolpert, D. (1992). Stacked generalization. *Neural Networks* 5(2), 241–259.
- Wolpert, D. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation* 8(7), 1341–1390.
- Wong, F., C. Carter, and R. Kohn (2003). Efficient estimation of covariance selection models. *Biometrika* 90(4), 809–830.
- Wood, F., C. Archambeau, J. Gasthaus, L. James, and Y. W. Teh (2009). A stochastic memoizer for sequence data. In *Intl. Conf. on Machine Learning*.
- Wright, S., R. Nowak, and M. Figueiredo (2009). Sparse reconstruction by separable approximation. *IEEE Trans. on Signal Processing* 57(7), 2479–2493.
- Wu, T. T. and K. Lange (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat* 2(1), 224–244.
- Wu, Y., H. Tjelmeland, and M. West (2007). Bayesian CART: Prior structure and MCMC computations. *J. of Computational and Graphical Statistics* 16(1), 44–66.
- Xu, F. and J. Tenenbaum (2007). Word learning as Bayesian inference. *Psychological Review* 114(2).
- Xu, Z., V. Tresp, A. Rettinger, and K. Kersting (2008). Social network mining with nonparametric relational models. In *ACM Workshop on Social Network Mining and Analysis (SNA-KDD 2008)*.
- Xu, Z., V. Tresp, K. Yu, and H.-P. Kriegel (2006). Infinite hidden relational models. In *UAI*.
- Xu, Z., V. Tresp, S. Yu, K. Yu, and H.-P. Kriegel (2007). Fast inference in infinite hidden relational models. In *Workshop on Mining and Learning with Graphs*.
- Xue, Y., X. Liao, L. Carin, and B. Krishnapuram (2007). Multi-task learning for classification with dirichlet process priors. *J. of Machine Learning Research* 8, 2007.
- Yadollahpour, P., D. Batra, and G. Shakhnarovich (2011). Diverse M-best Solutions in MRFs. In *NIPS workshop on Discrete Optimization in Machine Learning*.
- Yan, D., L. Huang, and M. I. Jordan (2009). Fast approximate spectral clustering. In *15th ACM Conf. on Knowledge Discovery and Data Mining*.
- Yang, A., A. Ganesh, S. Sastry, and Y. Ma (2010, Feb). Fast ℓ_1 -minimization algorithms and an application in robust face recognition: A review. Technical Report UCB/EECS-2010-13, EECS Department, University of California, Berkeley.

- Yang, C., R. Duraiswami, and L. David (2005). Efficient kernel machines using the improved fast Gauss transform. In *NIPS*.
- Yang, S., B. Long, A. Smola, H. Zha, and Z. Zheng (2011). Collaborative competitive filtering: learning recommender using context of user choice. In *Proc. Annual Intl. ACM SIGIR Conference*.
- Yanover, C., O. Schueler-Furman, and Y. Weiss (2007). Minimizing and Learning Energy Functions for Side-Chain Prediction. In *Recomb*.
- Yaun, G.-X., K.-W. Chang, C.-J. Hsieh, and C.-J. Lin (2010). A Comparison of Optimization Methods and Software for Large-scale L1-regularized Linear Classification. *J. of Machine Learning Research* 11, 3183–3234.
- Yedidia, J., W. T. Freeman, and Y. Weiss (2001). Understanding belief propagation and its generalizations. In *Intl. Joint Conf. on AI*.
- Yoshida, R. and M. West (2010). Bayesian learning in sparse graphical factor models via annealed entropy. *J. of Machine Learning Research* 11, 1771–1798.
- Younes, L. (1989). Parameter estimation for imperfectly observed Gibbsian fields. *Probab. Theory and Related Fields* 82, 625–645.
- Yu, C. and T. Joachims (2009). Learning structural SVMs with latent variables. In *Intl. Conf. on Machine Learning*.
- Yu, S., K. Yu, V. Tresp, K. H-P., and M. Wu (2006). Supervised probabilistic principal component analysis. In *Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining*.
- Yu, S.-Z. and H. Kobayashi (2006). Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden Markov model. *IEEE Trans. on Signal Processing* 54(5), 1947–1951.
- Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *J. Royal Statistical Society, Series B* 68(1), 49–67.
- Yuan, M. and Y. Lin (2007). Model selection and estimation in the gaussian graphical model. *Biometrika* 94(1), 19–35.
- Yuille, A. (2001). CCCP algorithms to minimize the Bethe and Kikuchi free energies: convergent alternatives to belief propagation. *Neural Computation* 14, 1691–1722.
- Yuille, A. and A. Rangarajan (2003). The concave-convex procedure. *Neural Computation* 15, 915.
- Yuille, A. and S. Zheng (2009). Compositional noisy-logical learning. In *Intl. Conf. on Machine Learning*.
- Yuille, A. L. and X. He (2011). Probabilistic models of vision and max-margin methods. *Frontiers of Electrical and Electronic Engineering* 7(1).
- Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. In *Bayesian inference and decision techniques, Studies of Bayesian and Econometrics and Statistics volume 6*. North Holland.
- Zhai, C. and J. Lafferty (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Trans. on Information Systems* 22(2), 179–214.
- Zhang, N. (2004). Hierarchical latent class models for cluster analysis. *J. of Machine Learning Research*, 301–308.
- Zhang, N. and D. Poole (1996). Exploiting causal independence in Bayesian network inference. *J. of AI Research*, 301–328.
- Zhang, T. (2008). Adaptive Forward-Backward Greedy Algorithm for Sparse Learning with Linear Models. In *NIPS*.
- Zhang, X., T. Graepel, and R. Herbrich (2010). Bayesian Online Learning for Multi-label and Multi-variate Performance Measures. In *AI/Statistics*.
- Zhao, J.-H. and P. L. H. Yu (2008, November). Fast ML Estimation for the Mixture of Factor Analyzers via an ECM Algorithm. *IEEE Trans. on Neural Networks* 19(11).
- Zhao, P., G. Rocha, and B. Yu (2005). Grouped and Hierarchical Model Selection through Composite Absolute Penalties. Technical report, UC Berkeley.
- Zhao, P. and B. Yu (2007). Stagewise Lasso. *J. of Machine Learning Research* 8, 2701–2726.
- Zhou, H., D. Karakos, S. Khudanpur, A. Andreou, and C. Priebe (2009). On Projections of Gaussian Distributions using Maximum Likelihood Criteria. In *Proc. of the Workshop on Information Theory and its Applications*.
- Zhou, M., H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin (2009). Non-parametric Bayesian Dictionary Learning for Sparse Image Representations. In *NIPS*.
- Zhou, X. and X. Liu (2008). The EM algorithm for the extended finite mixture of the factor analyzers model. *Computational Statistics and Data Analysis* 52, 3939–3953.
- Zhu, C. S., N. Y. Wu, and D. Mumford (1997, November). Minimax entropy principle and its application to texture modeling. *Neural Computation* 9(8).
- Zhu, J. and E. Xing (2010). Conditional topic random fields. In *Intl. Conf. on Machine Learning*.
- Zhu, L., Y. Chen, A. Yuille, and W. Freeman (2010). Latent hierarchical structure learning for object detection. In *CVPR*.
- Zhu, M. and A. Ghodsi (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis* 51, 918–930.
- Zhu, M. and A. Lu (2004). The counterintuitive non-informative prior for the bernoulli family. *J. Statistics Education*.
- Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Intl. Conf. on Machine Learning*, pp. 928–936.
- Zobay, O. (2009). Mean field inference for the Dirichlet process mixture model. *Electronic J. of Statistics* 3, 507–545.
- Zoeter, O. (2007). Bayesian generalized linear models in a terabyte world. In *Proc. 5th International Symposium on Image and Signal Processing and Analysis*.

- Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. of the Am. Stat. Assoc.*, 1418–1429.
- Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *J. of Royal Stat. Soc. Series B* 67(2), 301–320.
- Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *J. of Computational and Graphical Statistics* 15(2), 262–286.
- Zou, H., T. Hastie, and R. Tibshirani (2007). On the "Degrees of Freedom" of the Lasso. *Annals of Statistics* 35(5), 2173–2192.
- Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* 36(4), 1509–1533.
- Zweig, G. and M. Padmanabhan (2000). Exact alpha-beta computation in logarithmic space with application to map word graph construction. In *Proc. Intl. Conf. Spoken Lang.*

Index to code

agglomDemo, 894
amazonSellerDemo, 155
arsDemo, 819
arsEnvelope, 819

bayesChangeOfVar, 151
bayesLinRegDemo2d, 233
bayesTtestDemo, 138
beliefPropagation, 768
bernoulliEntropyFig, 57
besselk, 477
betaBinomPostPredDemo, 79
betaCredibleInt, 153
betaHPD, 153, 154
betaPlotDemo, 43
biasVarModelComplexity3, 204
bimodalDemo, 150
binaryFaDemoTipping, 403
binomDistPlot, 35
binomialBetaPosteriorDemo, 75
bleiLDAperplexityPlot, 955
bolassoDemo, 440
boostingDemo, 555, 558
bootstrapDemoBer, 192

cancerHighDimClassifDemo, 110
cancerRatesEb, 172
casinoDemo, 606, 607
centralLimitDemo, 52
changeOfVarsDemold, 53
chowliuTreeDemo, 913
coinsModelSelDemo, 164
contoursSSEdemo, 219
convexFnHand, 222
curseDimensionality, 18

demard, 580
depnetFit, 909
dirichlet3dPlot, 48
dirichletHistogramDemo, 48
discreteProbDistFig, 28
discrimAnalysisDboundariesDemo, 103, 105
discrimAnalysisFit, 106
discrimAnalysisHeightWeightDemo, 145
discrimAnalysisPredict, 106
dpmGauss2dDemo, 888
dpmSampleDemo, 881
dtfit, 545
dtreeDemoIris, 549, 550

elasticDistortionsDemo, 567
emLogLikelihoodMax, 365

faBiplotDemo, 383
fisherDiscrimVowelDemo, 274
fisheririsDemo, 6
fisherLDAdemo, 272
fmGibbs, 843

gammaPlotDemo, 41, 150
gammaRainfallDemo, 41
gampdf, 41
gaussCondition2Ddemo2, 112
gaussHeightWeight, 102
gaussImputationDemo, 115, 375

gaussInferParamsMean1d, 121
gaussInferParamsMean2d, 123
gaussInterpDemo, 113
gaussInterpNoisyDemo, 125
gaussMissingFitEm, 374
gaussMissingFitGibbs, 840
gaussPlot2d, 142
gaussPlot2Ddemo, 47
gaussPlotDemo, 19
gaussSeqUpdateSigma1D, 131
generativeVsDiscrim, 269
geomRidge, 229
ggmFitDemo, 939
ggmFitHtf, 939
ggmFitMinfunc, 939
ggmLassoDemo, 13, 940
ggmLassoHtf, 940
gibbsDemolsing, 670, 873
gibbsGaussDemo, 848
giniDemo, 548
gpcDemo2d, 529
gpnnDemo, 536
gprDemoArd, 520
gprDemoChangeHparams, 519
gprDemoMarglik, 522
gprDemoNoiseFree, 517
gpSpatialDemoLaplace, 532
groupLassoDemo, 451

hclustYeastDemo, 894, 896
hingeLossPlot, 211, 556
hmmFilter, 609
hmmFwdBack, 611
hmmLilypadDemo, 604
hmmSelfLoopDist, 623
hopfieldDemo, 670
huberLossDemo, 223, 497

icaBasisDemo, 471
icaDemo, 408
icaDemoUniform, 409
IPFDemo2x2, 683
isingImageDenoiseDemo, 739, 839

kalmanFilter, 641
kalmanTrackingDemo, 632
kernelBinaryClassifDemo, 489
kernelRegrDemo, 490, 491
kernelRegressionDemo, 510
KLfwdReverseMixGauss, 734
KLpqGauss, 734
kmeansHeightWeight, 10
kmeansModelSelld, 371
kmeansYeastDemo, 341
knnClassifyDemo, 17, 23–25
knnVoronoi, 16
kpcaDemo2, 495
kpcaScholkopf, 493

lassoPathProstate, 437, 438
LassoShooting, 441
leastSquaresProjection, 221
linregAllsubsetsGraycodeDemo, 423
linregBayesCaterpillar, 237, 238
linregCensoredSchmeeHahnDemo, 379

- linregDemol, 241
- linregEbModelSelVsN, 158, 159, 749
- linregFitLITest, 447
- linregOnlineDemoKalman, 636
- linregPolyLassoDemo, 436
- linregPolyVsDegree, 9, 20, 436
- linregPolyVsN, 231
- linregPolyVsRegDemo, 208, 225, 226, 239
- linregPostPredDemo, 235
- linregRbfDemo, 487
- linregRobustDemoCombined, 223
- linregWedgeDemo2, 19
- LMSdemo, 265
- logregFit, 254
- logregLaplaceGirolamiDemo, 257, 258
- logregMultinomKernelDemo, 269
- logregSATdemo, 21
- logregSATdemoBayes, 259
- logregSatMhDemo, 852
- logregXorDemo, 486
- logsumexp, 86
- lossFunctionFig, 179
- lsiCode, 419
- marsDemo, 554
 - mcAccuracyDemo, 55
 - mcEstimatePi, 54
 - mcmcGmmDemo, 851, 860, 861
 - mcQuantileDemo, 153
 - mcStatDist, 598
 - miMixedDemo, 59
 - mixBerMnistEM, 341
 - mixBetaDemo, 170
 - mixexpDemo, 343
 - mixexpDemoOneToMany, 344
 - mixGaussDemoFaithful, 353
 - mixGaussLikSurfaceDemo, 346
 - mixGaussMLvsMAP, 356
 - mixGaussOverRelaxedEmDemo, 369
 - mixGaussPlotDemo, 339
 - mixGaussSingularity, 356
 - mixGaussVbDemoFaithful, 753, 755
 - mixPpcaDemoNetlab, 386
 - mixStudentBankruptcyDemo, 361
 - mlpPriorsDemo, 574
 - mlpRegEvidenceDemo, 579
 - mlpRegHmcDemo, 579
 - mnistlNNDemo, 25, 1002
 - multilevelLinregDemo, 844
 - mutualInfoAllPairsMixed, 59
- naiveBayesBowDemo, 84, 88
 - naiveBayesFit, 83, 277
 - naiveBayesPredict, 86, 277
 - netflixResultsPlot, 981
 - newsgroupsVisualize, 5
 - newtonsMethodMinQuad, 250
 - newtonsMethodNonConvex, 250
 - ngramPlot, 592
 - NIXdemo2, 135
 - normalGammaPenaltyPlotDemo, 460
 - normalGammaThresholdPlotDemo, 461
 - numbersGame, 69–71
- pagerankDemo, 600, 603
 - pagerankDemoPmtk, 602
 - paretoPlot, 44
 - parzenWindowDemo2, 509
 - pcaDemo2d, 388
 - pcaDemo3d, 11
 - pcaDemoHeightWeight, 389
 - pcaEmStepByStep, 397
 - pcaImageDemo, 12, 389
 - pcaOverfitDemo, 400–402
 - pcaPmtk, 393
 - pfColorTrackerDemo, 830
 - poissonPlotDemo, 37
 - postDensityIntervals, 154
 - ppcaDemo2d, 388
 - PRhand, 182
 - probitPlot, 259
 - probitRegDemo, 259, 294, 363
 - prostateComparison, 436
 - prostateSubsets, 427
- quantileDemo, 33
- randomWalk0to20Demo, 856
 - rbpfManeuverDemo, 834, 835
 - rbpfSlamDemo, 835
 - rdaFit, 108
 - regtreeSurfaceDemo, 545
 - rejectionSamplingDemo, 818
 - relevanceNetworkNewsgroupDemo, 908
 - residualsDemo, 219
 - ridgePathProstate, 437
 - riskFnGauss, 198
 - robustDemo, 40
 - robustPriorDemo, 168
- saDemoPeaks, 869, 870
 - sampleCdf, 816
 - samplingDistGaussShrinkage, 203
 - sensorFusion2d, 123
 - sensorFusionUnknownPrec, 141
 - seqlogoDemo, 36
 - shrinkageDemoBaseball, 175
 - shrinkcov, 130
 - shrinkcovDemo, 129
 - shrunkCentroidsFit, 109
 - shrunkCentroidsSRBCTdemo, 109, 110
 - shuffledDigitsDemo, 7, 25
 - sigmoidLowerBounds, 761
 - sigmoidPlot, 21
 - sigmoidplot2D, 246
 - simpsonsParadoxGraph, 933
 - sliceSamplingDemo1d, 865
 - sliceSamplingDemo2d, 865
 - smoothingKernelPlot, 507
 - softmaxDemo2, 103
 - SpaRSA, 445
 - sparseDictDemo, 471
 - sparseNnetDemo, 575
 - sparsePostPlot, 459
 - sparseSensingDemo, 438
 - spectralClusteringDemo, 893
 - splineBasisDemo, 125
 - ssmTimeSeriesSimple, 638, 639
 - steepestDescentDemo, 247, 248
 - stickBreakingDemo, 883
 - studentLaplacePdfPlot, 40
 - subgradientPlot, 432
 - subSuperGaussPlot, 412
 - surfaceFitDemo, 218
 - svdImageDemo, 394
 - svmCgammaDemo, 504
- tanhPlot, 570
 - trueskillDemo, 798

trueskillPlot, 797

unigaussVbDemo, 745

varEMbound, 368

variableElimination, 717

visDirichletGui, 48

visualizeAlarmNetwork, 314

vqDemo, 354

wiPlotDemo, 127

Index to keywords

- #P-hard, 727
- 0-1 loss, **177**
- 3-SAT, 727

- A star search, 887
- absorbing state, **598**
- accept, **848**
- action, 176
- action nodes, **328**
- action space, **176**
- actions, 176
- activation, **563**
- active learning, 230, 234, 938
- Active set, **441**
- active set, **442**
- Activity recognition, **605**
- Adaboost.M1, **559**
- adagrad, **263**
- adaline, **569**
- adaptive basis-function model, **543**
- adaptive importance sampling, **821**
- adaptive lasso, **460**
- adaptive MCMC, **853**
- adaptive rejection Metropolis sampling, **820**
- adaptive rejection sampling, **820**
- add-one smoothing, 77, 593
- ADF, **653**, 983
- adjacency matrix, **309**, **970**
- adjust for, **934**
- adjusted Rand index, **878**
- admissible, **197**
- admixture mixture, **950**
- AdSense, 928
- AdWords, 928
- affinity propagation, **887**
- agglomerative clustering, **893**
- agglomerative hierarchical clustering, 927
- aha, **68**
- AI, 1007
- AIC, **162**, 557
- Akaike information criterion, **162**
- alarm network, **313**
- alignment, 701
- all pairs, **503**
- alleles, **317**
- alpha divergence, **735**
- alpha expansion, **803**
- alpha-beta swap, **804**
- alternative hypothesis, **163**
- analysis view, **390**
- analysis-synthesis, **470**
- ancestors, **309**
- ancestral graph, **664**
- ancestral sampling, **822**
- and-or graphs, 1007
- annealed importance sampling, **871**, 923
- annealing, 853
- annealing importance sampling, 992
- ANOVA, 553
- anti-ferromagnets, **668**
- aperiodic, **598**
- approximate inference, **727**
- approximation error, **230**
- ARD, 238, **463**, 520, 580
- ARD kernel, **480**
- area under the curve, **181**

- ARMA, **639**, 674
- array CGH, 454
- association rules, 15
- associative, **931**
- associative Markov network, **668**
- associative memory, **568**, **669**, 997
- associative MRF, **802**
- assumed density filter, 267
- assumed density filtering, **653**, 787
- asymptotically normal, **194**
- asymptotically optimal, **201**
- asynchronous updates, **774**
- atom, **469**
- atomic bomb, 52
- attractive MRF, **802**
- attributes, 2, 3
- AUC, **181**
- audio-visual speech recognition, **628**
- augmented DAG, **932**
- auto-encoder, **1000**
- auto-encoders, 990
- auto-regressive HMM, **626**
- autoclass, **11**
- autocorrelation function, **862**
- automatic relevance determination, **463**
- automatic relevancy determination, 238, 398, 580, 747
- Automatic speech recognition, **605**
- automatic speech recognition, 624
- auxiliary function, **350**
- auxiliary variables, **863**, 868
- average link clustering, **897**
- average precision, 303
- average precision at K, **183**
- axis aligned, **47**
- axis parallel splits, **544**

- back-propagation, 999
- backdoor path, **934**
- backfitting, **552**, 563, **998**
- background knowledge, 68
- backoff smoothing, **594**
- backpropagation, **570**, 970
- backpropagation algorithm, 569
- backslash operator, **228**
- Backwards selection, **428**
- bag of words, **5**, **81**, **945**
- bag-of-characters, 483
- bag-of-words, 483
- bagging, **551**
- bandwidth, **480**, **507**
- barren node removal, 334, **714**
- BART, **551**, 586
- Barzilai-Borwein, **445**
- base distribution, **338**
- base learner, **554**
- base measure, **882**
- base rate fallacy, **30**
- basic feasible solution, 468
- basis function expansion, **20**, **217**
- basis functions, 421
- basis pursuit denoising, **430**
- batch, **261**
- Baum-Welch, **618**
- Bayes ball algorithm, **324**
- Bayes decision rule, **177**, **195**
- Bayes estimator, **177**, **195**

- Bayes factor, 137, **163**, 921
- Bayes model averaging, **71**, **581**
- Bayes point, **257**
- Bayes risk, **195**
- Bayes rule, **29**, 340
- Bayes Theorem, **29**
- Bayesian, xxvii, **27**
- Bayesian adaptive regression trees, **551**
- Bayesian factor regression, **405**
- Bayesian hierarchical clustering, **899**
- Bayesian information criterion, **161**
- Bayesian IPF, 683
- Bayesian lasso, **448**
- Bayesian model selection, **156**
- Bayesian network structure learning, **914**
- Bayesian networks, **310**
- Bayesian Occam's razor, **156**
- Bayesian statistics, **149**, 191
- BDe, **917**
- BDeu, **918**
- beam search, 428, 887
- belief networks, **310**
- belief propagation, 611, **707**, 767
- belief state, **71**, **332**, **607**, **609**
- belief state MDP, **332**
- belief updating, **709**
- bell curve, **20**, **38**
- Berkson's paradox, **326**
- Bernoulli, **21**, **34**
- Bernoulli product model, **88**
- Bernoulli-Gaussian, **426**
- Bessel function, 483
- beta distribution, **42**, 74
- beta function, 42
- beta process, 470
- beta-binomial, **78**
- Bethe, **781**
- Bethe energy functional, **781**
- Bethe free energy, **781**
- BFGS, **251**
- Bhattacharya distance, **828**
- bi-directed graph, **674**
- bias, **20**, **200**, 457
- bias term, **669**
- bias-variance tradeoff, **202**
- BIC, **161**, 256, 557, 920
- biclustering, **903**
- big data, 1
- bigram model, **591**
- binary classification, **3**, 65
- binary entropy function, **57**
- binary independence model, **88**
- binary mask, **426**, 470
- binary tree, **895**
- Bing, 302, 799, 983
- binomial, **34**
- binomial coefficient, **34**
- binomial distribution, 74
- binomial regression, 292
- BinomialBoost, **561**
- BIO, **687**
- biosequence analysis, **36**, 170
- bipartite graph, **313**
- biplot, **383**
- birth moves, 370
- bisecting K-means, **898**
- bits, **56**
- bits-back, 733
- black swan paradox, **77**, 84
- black-box, **340**, **585**
- Blackwell-MacQueen, **884**
- blank slate, **165**
- blind signal separation, **407**
- blind source separation, **407**
- blocked Gibbs sampling, **848**
- blocking Gibbs sampling, **848**
- bloodtype, 317
- BN2O, **315**
- bolasso, **439**
- Boltzmann distribution, **104**, **869**
- Boltzmann machine, 568, **669**, **983**
- bond variables, **866**
- Boosting, **554**
- boosting, 553, 742
- bootstrap, **192**
- bootstrap filter, **827**
- bootstrap lasso, **439**
- bootstrap resampling, 439
- borrow statistical strength, **171**, 231, 296, 845
- bottleneck, 205, **337**, **1000**
- bottleneck layer, **970**
- bound optimization, **369**
- box constraints, 444
- Box-Muller, **817**
- boxcar kernel, 508, **508**
- Boyen-Koller, **654**
- BP, **707**
- BPDN, **430**
- Bradley Terry, 795
- branch and bound, 811
- branching factor, **954**
- bridge regression, **458**
- Brownian motion, 483
- bucket elimination, **715**
- BUGS, 756, **847**
- Buried Markov models, **627**
- burn-in phase, **856**
- burned in, **838**
- burstiness, **88**
- bursty, 480
- C4.5, **545**
- calculus of variations, 289
- calibration, **724**
- Candidate method, **872**
- Canonical correlation analysis, **407**
- canonical form, **282**
- canonical link function, **291**
- canonical parameters, **115**, **282**
- Cardinality constraints, 810
- CART, **544**, 545
- Cartesian, 51
- cascade, **776**
- case analysis, **260**
- categorical, **2**, **35**
- categorical PCA, **402**, **947**, 961
- categorical variables, 876
- Cauchy, **40**
- causal Markov assumption, **931**
- Causal models, **931**
- causal MRF, **661**
- causal networks, **310**
- causal sufficiency, **931**
- causality, 919, 929
- CCA, **407**
- CCCP, **702**
- CD, **989**
- cdf, **32**, 38
- Censored regression, **379**
- censored regression, 380

- centering matrix, **494**
- central composite design, **523**
- central interval, **152**
- central limit theorem, **38, 51, 255**
- central moment, **413**
- central-limit theorem, **55**
- centroid, **341**
- centroids, **486**
- certainty factors, **675**
- chain graph, **671**
- chain rule, **29, 307**
- chance nodes, **328**
- change of variables, **50**
- channel coding, **56**
- Chapman-Kolmogorov, **590**
- characteristic length scale, **480**
- Cheeseman-Stutz approximation, **923**
- Chi-squared distribution, **42**
- chi-squared statistic, **163, 213**
- children, **309, 310**
- Chinese restaurant process, **884**
- chip-Seq, **622**
- Cholesky decomposition, **227, 817**
- Chomsky normal form, **689**
- chordal, **665**
- chordal graph, **720**
- Chow-Liu algorithm, **312, 912**
- CI, **308**
- circuit complexity, **944**
- city block distance, **876**
- clamped phase, **987**
- clamped term, **677**
- clamping, **319**
- class imbalance, **503**
- class-conditional density, **30, 65**
- classical, **149**
- classical statistics, **191**
- classification, **2, 3**
- Classification and regression trees, **544**
- clausal form, **675**
- clause, **727**
- click-through rate, **4**
- clique, **310**
- cliques, **719, 722**
- closing the loop, **635**
- closure, **662**
- cluster variational method, **783**
- Clustering, **875**
- clustering, **10, 340**
- clusters, **487**
- clutter problem, **788**
- co-clustering, **979**
- co-occurrence matrix, **5**
- co-parents, **327**
- coarse-to-fine grid, **775**
- cocktail party problem, **407**
- coclustering, **903**
- codebook, **354**
- collaborative filtering, **14, 300, 387, 903, 979**
- collapsed Gibbs sampler, **841**
- collapsed Gibbs sampling, **956**
- collapsed particles, **831**
- collect evidence, **707**
- collect-to-root, **723**
- collider, **324**
- COLT, **210**
- committee method, **580**
- commutative semi-ring, **717**
- commutative semiring, **726**
- compactness, **897**
- compelled edges, **915**
- complementary prior, **997**
- complete, **322**
- complete data, **270, 349**
- complete data assumption, **914**
- complete data log likelihood, **348, 350**
- complete link clustering, **897**
- completing the square, **143**
- composite likelihood, **678**
- compressed sensing, **472**
- compressive sensing, **472**
- computation tree, **772**
- computational learning theory, **210**
- computationalism, **569**
- concave, **222, 286**
- concave-convex procedure, **702**
- concentration matrix, **46**
- concentration parameter, **882**
- concept, **65**
- concept learning, **65**
- condensation, **827**
- conditional entropy, **59**
- conditional Gamma Poisson, **949**
- conditional Gaussian, **920**
- conditional independence, **308**
- conditional likelihood, **620**
- conditional logit model, **252**
- conditional probability, **29**
- conditional probability distribution, **308**
- conditional probability tables, **308**
- conditional random field, **684**
- conditional random fields, **606, 661**
- conditional topic random field, **969**
- conditionally conjugate, **132**
- conditionally independent, **31, 82**
- conditioning, **319**
- conditioning case, **322**
- conductance, **858**
- confidence interval, **212**
- confidence intervals, **153**
- confounder, **674**
- confounders, **931**
- confounding variable, **934**
- confusion matrix, **181**
- conjoint analysis, **297**
- conjugate gradients, **249, 524**
- conjugate prior, **74**
- conjugate priors, **281, 287**
- conjunctive normal form, **675**
- connectionism, **569**
- consensus sequence, **36, 606**
- conservation of probability mass, **157**
- consistent, **200**
- consistent estimator, **233**
- consistent estimators, **70**
- constant symbols, **676**
- constraint satisfaction problems, **717, 726**
- constraint-based approach, **924**
- content addressable memory, **669**
- context free grammar, **689**
- context specific independence, **321**
- context-specific independence, **944**
- contextual bandit, **184, 254**
- contingency table, **682**
- continuation method, **442, 869**
- contrastive divergence, **569, 989**
- contrastive term, **677**
- control signal, **625, 631**
- converge, **857**
- convex, **58, 221, 247, 285, 677**

- convex belief propagation, **785**, 943
- convex combination, 76, 130, **338**
- convex hull, **777**
- convolutional DBNs, **1004**
- convolutional neural nets, 1004
- convolutional neural network, **565**
- cooling schedule, **870**
- corpus, **953**
- correlated topic model, 757, **961**
- correlation coefficient, **45**, 876
- correlation matrix, **45**
- correspondence, **658**
- cosine similarity, **480**
- cost-benefit analysis, **186**
- coupled HMM, **628**
- covariance, **44**
- covariance graph, **674**, 908
- covariance matrix, **45**, 46
- covariance selection, **938**
- covariates, **2**
- CPD, **308**
- CPTs, **308**
- Cramer-Rao inequality, 201
- Cramer-Rao lower bound, **201**
- credible interval, **137**, **152**, 212
- CRF, 661, **684**
- critical temperature, **868**
- critical value, **671**
- cross entropy, **57**, 571
- cross over rate, **181**
- cross validation, **24**, **206**
- cross-entropy, **246**, 953
- cross-language information retrieval, 963
- crosscat, **904**
- crowd sourcing, **10**, 995
- CRP, **884**
- CTR, **4**
- cubic spline, **537**
- cumulant function, **282**, **284**
- cumulants, **284**
- cumulative distribution function, **32**, 38
- curse of dimensionality, **18**, 487
- curved exponential family, **282**
- cutting plane, **698**
- CV, **24**
- cycle, **310**
- cyclic permutation property, **99**
- d-prime, **106**
- d-separated, **324**
- DACE, 518
- DAG, **310**
- damped updates, **739**
- damping, **773**
- Dasher, 591
- data association, **658**, 810
- data augmentation, 362, **847**
- data compression, **56**
- data fragmentation, **546**
- data fusion, **404**
- data overwhelms the prior, **69**
- data-driven MCMC, **853**
- data-driven proposals, **828**
- DBM, **996**
- DBN, **628**, **997**
- DCM, **89**
- DCT, 469
- death moves, 370
- debiasing, **439**
- decision, 176
- decision boundary, 22
- decision diagram, **328**
- decision nodes, **328**
- decision problem, 176
- decision procedure, **177**
- decision rule, **22**
- decision trees, **544**
- decoding, **693**
- decomposable, **665**, **722**, 941
- decomposable graphs, 682
- decomposes, **322**, **917**
- DeeBN, **628**
- DeeBNs, **997**
- deep, 929
- deep auto-encoders, **1000**
- deep belief network, **997**
- deep Boltzmann machine, **996**
- deep directed networks, **996**
- deep learning, 479, **995**
- deep networks, 569
- defender's fallacy, **61**
- deflated matrix, **418**
- degeneracy problem, **825**
- degenerate, **532**, 535
- degree, **310**
- degrees of freedom, **39**, **161**, 206, **229**, **534**
- deleted interpolation, **593**
- delta rule, **265**
- dendrogram, **895**
- denoising auto-encoder, **1001**
- dense stereo reconstruction, **690**
- density estimation, **9**
- dependency network, **909**
- dependency networks, 679
- derivative free filter, **651**
- descendants, **309**
- descriptive, **2**
- design matrix, 3, 875
- detailed balance, **854**
- detailed balance equations, **599**
- determinism, 944
- deterministic annealing, **367**, 620
- deviance, **547**
- DGM, **310**
- diagonal, 46
- diagonal covariance LDA, 107
- diagonal LDA, **108**
- diameter, **710**, **897**
- dictionary, **469**
- digamma, 361, 752, 958
- digital cameras, 8
- dimensionality reduction, **11**, 1000
- Dirac delta function, **39**
- Dirac measure, **37**, **68**
- Dirichlet process, 903
- direct posterior probability approach, **184**
- directed, **309**
- directed acyclic graph, **310**
- directed graphical model, **310**
- directed local Markov property, **327**
- directed mixed graph, 929
- directed mixed graphical model, **674**
- Dirichlet, 79
- Dirichlet Compound Multinomial, **89**
- Dirichlet distribution, **47**
- Dirichlet multinomial regression LDA, **969**
- Dirichlet process, 596, **879**, **882**, 973, 976
- Dirichlet process mixture models, 508, 755
- discontinuity preserving, **691**
- discounted cumulative gain, **303**

- discrete, **35**
- discrete AdaBoost, **559**
- discrete choice modeling, **296**
- discrete random variable, **28**
- discrete with probability one, **884**
- discretize, **59**, **691**
- discriminability, **106**
- discriminant analysis, **101**
- discriminant function, **500**
- discriminative, **245**
- discriminative classifier, **30**
- discriminative LDA, **968**
- discriminative random field, **684**
- disease mapping, **531**
- disease transmission, **970**
- disparity, **691**
- dispersion parameter, **290**
- dissimilarity analysis, **898**
- dissimilarity matrix, **875**
- distance matrix, **875**
- distance transform, **775**
- distorted, **566**
- distortion, **354**
- distribute evidence, **707**
- distribute-from-root, **724**
- distributed encoding, **984**
- distributed representation, **569**, **627**
- distributional particles, **831**
- distributive law, **717**
- divisive clustering, **893**
- DNA sequences, **36**
- do calculus, **932**
- Document classification, **87**
- document classification, **5**
- Domain adaptation, **297**
- domain adaptation, **297**
- dominates, **197**
- double loop algorithms, **773**
- double Pareto distribution, **461**
- double sided exponential, **41**
- dRUM, **294**
- dual decomposition, **808**
- dual variables, **492**, **499**
- dummy encoding, **35**
- dyadic, **976**
- DyBN, **628**
- DyBNs, **997**
- dynamic Bayes net, **653**
- dynamic Bayesian network, **628**
- dynamic linear model, **636**
- dynamic programming, **331**, **920**
- dynamic topic model, **962**

- E step, **350**
- e-commerce, **11**
- early stopping, **263**, **557**, **572**
- EB, **173**
- ECM, **369**, **387**
- ECME, **369**
- ECOC, **581**
- econometric forecasting, **660**
- economy sized SVD, **392**
- edge appearance probability, **786**
- edges, **309**
- edit distance, **479**
- EER, **181**
- effective sample size, **75**, **825**, **862**
- efficient IPF, **683**
- efficiently PAC-learnable, **210**
- eigendecomposition, **98**
- eigenfaces, **12**
- eigengap, **857**
- eigenvalue spectrum, **130**
- EKF, **648**
- elastic net, **438**, **456**, **936**
- elimination order, **718**
- EM, **271**, **349**, **618**, **749**
- email spam filtering, **5**
- embedding, **575**
- empirical Bayes, **157**, **162**, **173**, **300**, **746**
- empirical distribution, **37**, **205**
- empirical measure, **37**
- empirical risk, **205**, **697**
- empirical risk minimization, **205**, **261**
- end effector, **344**
- energy based models, **666**
- energy function, **255**
- energy functional, **732**, **778**
- ensemble, **980**
- Ensemble learning, **580**
- ensemble learning, **742**
- entanglement, **629**
- entanglement problem, **635**, **653**
- Entropy, **547**
- entropy, **56**
- EP, **983**
- Epanechnikov kernel, **508**
- ePCA, **947**
- epigraph, **222**
- epistemological uncertainty, **973**
- epoch, **264**, **566**
- epsilon insensitive loss function, **497**
- EPSR, **859**
- equal error rate, **181**
- equilibrium distribution, **597**
- equivalence class, **915**
- equivalent kernel, **512**, **533**
- equivalent sample size, **76**, **917**
- erf, **38**
- ergodic, **599**
- Erlang distribution, **42**
- ERM, **205**, **261**
- error bar, **76**
- error correcting codes, **768**
- error correction, **56**
- error function, **38**
- error signal, **265**
- error-correcting output codes, **503**, **581**
- ESS, **862**
- essential graph, **915**
- estimated potential scale reduction, **859**
- estimator, **191**
- Euclidean distance, **18**
- evidence, **156**, **173**
- evidence procedure, **173**, **238**, **746**
- evolutionary MCMC, **429**
- exchangeable, **321**, **963**
- exclusive or, **486**
- expectation correction, **658**
- expectation maximization, **349**
- expectation proagation, **735**
- Expectation propagation, **787**
- expectation propagation, **525**
- expected complete data log likelihood, **350**, **351**
- expected profit, **330**
- expected sufficient statistics, **350**, **359**, **619**
- expected value, **33**
- explaining away, **326**
- explicit duration HMM, **622**
- exploration-exploitation, **184**

- exploratory data analysis, **7**
- exponential cooling schedule, **870**
- Exponential distribution, **42**
- exponential family, 115, 253, **281, 282**, 290, 347
- exponential family harmonium, **985**
- exponential family PCA, **947**
- exponential loss, **556**
- exponential power distribution, **458**
- extended Kalman filter, **648**
- extension, **67**
- external field, **668**
- F score, **183**
- FI score, **183**, 699
- FA, **381**
- face detection, **8**
- face detector, 555
- face recognition, **8**
- Facebook, 974
- factor, **665**
- factor analysis, **381**, 402, 931, 947
- factor analysis distance, **520**
- factor graph, 769, **769**, 771, 888
- factor loading matrix, **381**
- factorial HMM, **628**
- factorial prior, **463**
- factors, **382**
- faithful, **936**
- false alarm, **30, 180**
- false alarm rate, **181**
- false discovery rate, **184**
- false negative, **180**
- false positive, **30, 180**
- false positive rate, **181**
- family, **309**
- family marginal, **359**
- fan-in, **313**
- fantasy data, **990**
- farthest point clustering, **355**
- fast Fourier transform, 717, 775
- fast Gauss transform, 524
- fast ICA, **411**
- fast iterative shrinkage thresholding algorithm, **446**
- FastSLAM, 635, **835**
- fat hand, **933**
- fault diagnosis, **659**
- feature construction, **564**
- feature extraction, **6, 564**
- feature function, **667**
- feature induction, **680**
- feature maps, **565**
- feature matrix, 875
- feature selection, **86**
- feature-based clustering, **875**
- features, 2, 3, **412**
- feedback loops, **929**
- feedforward neural network, **563**
- ferro-magnets, **668**
- FFT, 775
- fields of experts, **473**
- fill-in edges, **719**
- Filtering, **607**
- filtering, **87**
- finite difference matrix, **113**
- finite mixture model, 879
- first-order logic, **674**
- Fisher information, **166**
- Fisher information matrix, 152, **193**, 293
- Fisher kernel, **485**
- Fisher scoring method, **293**
- Fisher's linear discriminant analysis, **271**
- FISTA, **446**
- fit-predict cycle, **206**
- fixed effect, **298**
- Fixed lag smoothing, **608**
- fixed point, 139
- flat clustering, **875**
- FLDA, **271**
- flow cytometry, 936
- folds, **24**
- forest, **310, 912**
- forward stagewise additive modeling, **557**
- forward stagewise linear regression, **562**
- forwards KL, 733
- forwards model, **345**
- forwards selection, **428**
- forwards-backwards, 644, 688, 707, 720
- forwards-backwards algorithm, **428, 611**
- founder model, **317**
- founder variables, **385**
- Fourier basis, 472
- fraction of variance explained, **400**
- free energy, **988**
- free-form optimization, **737**
- frequent itemset mining, **15**
- frequentist, **27, 149**
- frequentist statistics, **191**
- Frobenius norm, **388**
- frustrated, **868**
- frustrated system, **668**
- full, 46
- full conditional, **328, 838**
- function approximation, **3**
- functional data analysis, **124**
- functional gradient descent, **561**
- furthest neighbor clustering, **897**
- fused lasso, **454**
- fuzzy clustering, **973**
- fuzzy set theory, **65**
- g-prior, **236**, 425
- game against nature, 176
- game theory, 176
- Gamma, 623
- gamma distribution, **41**
- gamma function, 42
- GaP, **949**
- gap statistic, **372**
- gating function, **342**
- Gauss-Seidel, 710
- Gaussian, **20, 38**
- Gaussian approximation, **255**, 731
- Gaussian Bayes net, **318**
- Gaussian copulas, **942**
- Gaussian graphical models, 725
- Gaussian kernel, **480, 507**, 517
- Gaussian mixture model, **339**
- Gaussian MRF, **672**
- Gaussian process, 483, 505, 509, 512, 882
- Gaussian processes, **515**
- Gaussian random fields, **938**
- Gaussian RBM, **986**
- Gaussian scale mixture, **359, 447**, 505
- Gaussian sum filter, **656**
- GDA, 101
- GEE, **300**
- GEM, **369**
- Gene finding, **606**
- gene finding, 622
- gene knockout experiment, 931

- gene microarrays, 421
- generalization, 3
- generalization error, **23**, **180**
- generalization gradient, **66**
- generalize, 3
- generalized additive model, **552**
- generalized belief propagation, **785**
- generalized cross validation, **207**
- generalized eigenvalue, **274**
- generalized EM, **361**, **369**
- generalized estimating equations, **300**
- generalized linear mixed effects model, **298**
- generalized linear model, 281, **290**
- generalized linear models, 281
- generalized pseudo Bayes filter, **657**
- generalized t distribution, **461**
- generate and test, **853**
- generative approach, 245
- generative classifier, **30**
- generative pre-training, **999**
- generative weights, **410**, **986**
- genetic algorithms, 348, 720, 921
- genetic linkage analysis, **315**, 318
- genome, 318
- genotype, **317**
- geometric distribution, **622**
- Gibbs distribution, **290**, **666**
- Gibbs sampler, 672
- Gibbs sampling, 328, 669, 736, **838**
- Gini index, **548**
- gist, 963
- Gittins Indices, **184**
- Glasso, **940**
- Glauber dynamics, **838**
- GLM, **290**, 654
- GLMM, **298**
- glmnet, 442
- global balance equations, **597**
- global convergence, **248**
- global localization, **828**
- global Markov property, **661**
- global minimum, 222
- global prior parameter independence, **916**
- globally normalized, **686**
- GM, **308**
- GMM, **339**
- GP-LVM, **540**
- GPs, **515**
- GPUs, 1006
- gradient boosting, **560**
- gradient descent, **247**, 445
- Gram matrix, **481**
- grammars, 689, 1007
- grandmother cells, **984**, 1005
- graph, **309**
- graph cuts, **890**
- graph Laplacian, **891**
- graph surgery, **932**
- graph-guided fused lasso, **454**
- graphcuts, **801**
- graphical lasso, **940**
- graphical model, **308**, 311
- graphical models, xxviii, 13, 31, 32, 308, 337, 909
- Gray code, **422**
- greatest common divisor, **598**
- greedy equivalence search, **936**
- ground network, **676**
- ground states, **668**
- group lasso, **450**, 579, 942
- grouping effect, **456**
- Gumbel, 295
- Hadamard product, **609**
- Haldane prior, **166**
- ham, 5
- Hamiltonian MCMC, **868**
- Hammersley-Clifford, **666**
- hamming distance, **876**
- handwriting recognition, 7
- haplotype, **317**
- hard clustering, **340**
- hard EM, **352**
- hard thresholding, **434**, 435
- harmonic mean, 183
- harmonium, **983**
- Hastings correction, **849**
- hat matrix, **221**
- HDI, **154**
- heat bath, **838**
- heavy ball method, **249**
- heavy tails, **43**, **223**
- Hellinger distance, **735**
- Helmholtz free energy, **733**
- Hessian, 193, 852
- heteroscedastic LDA, **275**
- heuristics, **727**
- hidden, **10**, 349
- hidden layer, **563**
- hidden Markov model, **312**, **603**, 963
- hidden nodes, **313**
- hidden semi-Markov model, **622**
- hidden units, **564**
- hidden variable, **312**, 924
- hidden variables, **319**, 914
- hierarchical adaptive lasso, **458**
- hierarchical Bayesian model, **171**
- hierarchical Bayesian models, 347
- hierarchical clustering, **875**, **893**
- hierarchical Dirichlet process, 621
- hierarchical HMM, **624**
- hierarchical latent class model, **926**
- hierarchical mixture of experts, **344**, 551
- high throughput, 184, 421
- high variance estimators, 550
- highest density interval, **154**
- highest posterior density, **153**
- hill climbing, 920
- hindsight, **607**
- hinge loss, **211**, 477, **499**
- Hinton diagram, 592
- Hinton diagrams, **399**
- histogram, **508**
- hit rate, **181**
- HMM, **312**, **603**
- HMM filter, 640
- HMMs, 685
- Hoeffding's inequality, **209**
- homogeneous, **589**
- homotopy, **442**
- Hopfield network, **568**, **669**
- horizon, **608**
- Horn clauses, **676**
- HPD, **153**
- HSMM, **622**
- Huber loss, **224**, 561
- Hugin, **722**
- Hungarian algorithm, **659**, 810
- hybrid MCMC, **868**
- hybrid Monte Carlo, 584
- hybrid systems, **655**

- hyper-parameters, **74**
- hypothesis space, **66**
- I-map, **324**
- I-projection, **733**
- ICA, **385, 409**
- ID3, **545**
- IDA, **936**
- identifiable, **346**
- identifiable in the limit, **70**
- iff, **68**
- iid, **51, 218, 320**
- ill-conditioned, **106, 129**
- image classification, **7**
- image compression, **355**
- image denoising, **473**
- image inpainting, **14, 473**
- image segmentation, **671**
- image tagging, **968**
- IMM, **658**
- implicit feedback, **983**
- importance sampling, **820**
- importance weights, **821**
- impression log, **983**
- improper prior, **166, 168**
- imputation, **14**
- Imputation Posterior, **847**
- in-degree, **310**
- inclusion probabilities, **423**
- incremental EM, **365, 366**
- independence sampler, **848**
- independent and identically distributed, **51**
- independent component analysis, **409**
- indicator function, **17, 28, 976**
- induced width, **719**
- induction, **66, 77**
- inductive bias, **19, 582**
- infer.net, **799**
- inference, **320**
- infinite hidden relational model, **977**
- infinite HMM, **621**
- infinite mixture models, **841, 879**
- infinite relational model, **903, 973, 976**
- influence diagram, **328, 932**
- influence model, **628**
- infomax, **416**
- information, **27**
- information arc, **329, 331**
- information bottleneck, **405**
- information extraction, **688**
- information filter, **642**
- information form, **115, 305, 672, 711, 725**
- information gain, **547**
- Information inequality, **58**
- information projection, **733**
- information retrieval, **183, 300, 953**
- information theory, **56**
- inheritance model, **317**
- inner approximation, **779**
- innovation, **641**
- inside outside, **624**
- inside-outside algorithm, **689**
- instance-based learning, **17**
- integrate out, **156**
- integrated likelihood, **156**
- integrated risk, **195**
- intensive care unit, **313**
- inter-causal reasoning, **326**
- interaction effects, **421**
- interactive multiple models, **658**
- interest point detector, **484**
- interpolate, **112**
- interpolated Kneser-Ney, **595**
- interpolator, **517**
- interval censored, **379**
- interventional data, **936**
- interventions, **931**
- intrinsic Gaussian random field, **113**
- invariant, **8, 854**
- invariant distribution, **597**
- invariant features, **1004**
- inverse chi-squared distribution, **131**
- inverse Gamma, **130**
- inverse gamma, **42**
- inverse Gaussian, **448**
- inverse probability transform, **815**
- inverse problem, **317**
- inverse problems, **344**
- inverse reinforcement learning, **186**
- inverse Wishart, **126, 128**
- inverted index, **600**
- inverted indices, **1004**
- IP, **847**
- IPF, **682**
- iris, **6, 548**
- IRLS, **251**
- IRM, **976**
- irreducible, **598**
- Ising model, **668**
- isotropic, **46**
- iterated EKF, **650**
- iterative conditional modes, **669, 804, 929**
- iterative proportional fitting, **682, 939**
- iterative scaling, **683**
- iterative shrinkage and thresholding algorithm, **445**
- iterative soft thresholding, **445**
- iteratively reweighted least squares, **251**
- Jacobi, **710, 773**
- Jacobian, **151, 648, 649**
- Jacobian matrix, **50**
- JAGS, **847**
- JamBayes, **13**
- James Stein estimator, **174**
- James-Stein estimator, **173, 199**
- JC Penney, **603**
- Jeffreys prior, **166**
- Jeffreys-Lindley paradox, **165**
- Jensen's inequality, **58, 363**
- Jensen-Shannon divergence, **57**
- Jeopardy, **4**
- jittered, **486**
- JJ bound, **761**
- joint distribution, **29, 307**
- joint probability distribution, **44**
- JTA, **720**
- jump Markov linear system, **655**
- junction tree, **722**
- junction tree algorithm, **720, 731**
- junction trees, **635**
- K-centers, **887**
- K-means algorithm, **352**
- k-means++, **355**
- K-medoids algorithm, **490**
- k-spectrum kernel, **484**
- K2 algorithm, **920**
- Kalman filter, **122, 267, 632, 633, 640, 643**
- Kalman gain matrix, **637, 641**
- Kalman smoother, **633, 707**

- Kalman smoothing, **644**, 712, 963
- Karhunen Loeve, **387**
- Karl Popper, 77
- KDE, **508**, 510
- Kendall's τ , **304**
- kernel, 565, **600**, **848**
- kernel density estimation, 127, 510
- kernel density estimator, **508**
- kernel function, **479**, 515
- kernel machine, **486**
- kernel PCA, **494**, 540, 892
- kernel regression, **511**
- kernel smoothing, **511**
- kernel trick, **488**
- kernelised feature vector, **486**
- Kikuchi free energy, **784**
- kinect, 551
- kinematic tracking, **344**
- kink, **372**
- KL divergence, **57**, 732
- Kleene star, 483
- knee, **372**
- KNN, **16**
- knots, **537**
- knowledge base, **676**
- knowledge discovery, **2**, **9**
- knowledge engineering, **313**
- Kolmogorov Smirnov, 864
- kriging, **516**
- kronecker product, **253**, 760
- Kruskal's algorithm, 912
- Kullback-Leibler divergence, **57**
- kurtosis, **413**, 415
- L-BFGS, **252**
- ℓ_0 pseudo-norm, 424
- ℓ_0 regularization, 426
- ℓ_1 loss, 179
- ℓ_1 regularization, 430
- L1-Adaboost, **563**
- L1VM, **488**, 505
- ℓ_2 loss, 179
- ℓ_2 norm, **218**
- ℓ_2 regularization, 226
- L2boosting, **558**
- L2VM, **488**
- label, 176
- label bias, **685**
- label switching, **341**, **841**
- label taxonomy, **689**
- labeled LDA, 953, **969**
- lag, 608
- Lagrange multiplier, **80**
- Lagrange multipliers, 289
- Lagrangian, **80**, 289
- Lagrangian relaxation, **808**
- Lanczos algorithm, 398
- language model, 300, **953**
- language modeling, **81**, 568
- language models, **591**
- Laplace, 223, 413, 429
- Laplace approximation, **255**, 468
- Laplace distribution, **41**
- Laplace's rule of succession, **77**
- LAR, **442**, 562
- large margin classifier, **501**
- large margin principle, 259
- LARS, **437**, **442**, 558, 562
- lasso, **431**, 470, 562, 936
- latent, **11**
- latent class model, **926**
- latent CRF, 701
- latent Dirichlet allocation, **949**, **950**
- latent factors, **11**
- latent semantic analysis, 12, **947**
- latent semantic indexing, **418**, **947**
- latent SVMs, **702**
- latent variable models, **337**
- lattice, 668
- Lauritzen-Spiegelhalter, **722**
- LBP, **767**
- LDA, **104**, 927, **949**, **950**
- LDA-HMM, **963**
- LDPC, 768
- LDS, **631**
- leaf, **309**
- leak node, **315**
- leapfrog steps, **868**
- learning, 320
- learning curve, **230**
- learning rate, **247**
- learning to learn, **296**
- learning to rank, **300**
- least favorable prior, **197**
- least mean squares, **265**, **637**
- least squares, **219**
- least squares boosting, 428, 442, **558**
- leave one out cross validation, **207**
- leave-one out cross validation, **24**
- leaves, 895
- left censored, **379**
- left-to-right, 612
- left-to-right transition matrix, **590**
- LeNet5, **566**
- leptokurtic, **413**
- LETOR, **300**
- level sets, 47
- Levenberg Marquardt, **250**
- Levinson-Durbin, 627
- LG-SSM, **631**
- likelihood, 319
- likelihood equivalence, **917**
- likelihood equivalent, 200
- likelihood principle, **214**
- likelihood ratio, **67**, **163**
- likelihood weighting, **822**
- limited memory BFGS, **252**
- limiting distribution, **598**
- line minimization, **248**
- line search, **248**
- linear discriminant analysis, **104**
- linear dynamical system, **631**
- linear Gaussian, **318**
- linear Gaussian system, **119**
- linear kernel, **482**
- linear program, **224**
- linear programming relaxation, **800**
- linear regression, **19**
- linear smoother, **533**
- linear threshold unit, **252**
- linear trend, 660
- linear-Gaussian CPD, 673
- linear-Gaussian SSM, **631**
- linearity of expectation, **49**
- linearly separable, **22**, 252, 266
- link farms, **601**
- link function, **291**
- LISREL, **930**
- ListNet, **302**
- LMS, **265**, **637**

- local consistency, **780**
- local evidence, **317, 671**
- local level model, **637**
- local prior parameter independence, **917**
- local variational approximation, **756**
- localist encoding, **984**
- locally decodable, **811**
- locally normalized, **686, 715**
- locally weighted regression, **512**
- LOESS, **512**
- log partition function, **282**
- log-linear, **667**
- log-loss, **210**
- log-odds ratio, **283**
- log-sum-exp, **86, 757**
- logic sampling, **822**
- logical reasoning problems, **726**
- logistic, **21, 295**
- logistic distribution, **413, 863**
- logistic normal, **402, 961**
- logistic regression, **21, 106**
- logit, **21**
- logitBoost, **560**
- long tail, **2, 296**
- long tails, **43**
- LOOCV, **24, 207**
- look-ahead RBPF, **832**
- loop, **310**
- loopy belief propagation, **691, 767, 889**
- Lorentz, **40**
- loss, **176**
- loss function, **261**
- loss matrix, **185**
- loss-augmented decoding, **699**
- loss-calibrated inference, **694**
- lossy compression, **354**
- low density parity check, **768**
- Low-level vision, **690**
- LOWESS, **512**
- LSA, **947, 1003**
- lse, **757**
- LSI, **947**
- LVM, **337**

- M step, **350**
- M-projection, **733**
- M3nets, **693**
- machine learning, **1**
- macro-averaged F1, **183**
- Mahalanobis distance, **98**
- mammogram, **29**
- maneuvering target tracking, **832**
- manifest, **930**
- MAP estimate, **4, 178**
- MAR, **270**
- margin, **563**
- margin re-rescaling, **696**
- marginal distribution, **29**
- marginal likelihood, **156, 169**
- marginal polytope, **777**
- marginalizing out, **320**
- marginally independent, **30**
- marker, **317**
- market basket analysis, **15**
- Markov, **324**
- Markov assumption, **308**
- Markov blanket, **327, 662, 736, 838**
- Markov chain, **308, 589**
- Markov Chain Monte Carlo, **815**
- Markov chain Monte Carlo, **52, 600, 837**
- Markov decision process, **331**
- Markov equivalence, **936**
- Markov equivalent, **915, 917**
- Markov logic network, **675**
- Markov mesh, **661**
- Markov model, **589**
- Markov models, **32**
- Markov network, **661**
- Markov random field, **661**
- Markov switching models, **604**
- MARS, **538, 553, 562**
- MART, **562**
- master, **810**
- matching pursuit, **562**
- matching pursuits, **428**
- Matern kernel, **482**
- MATLAB, **xxviii**
- matrix completion, **14, 939**
- matrix determinant lemma, **118**
- matrix factorization, **948**
- matrix inversion lemma, **118, 144, 641**
- matrix permanent, **669**
- matrix tree theorem, **914**
- max flow/min cut, **801**
- max margin Markov networks, **693**
- max pooling, **1005**
- max product linear programming, **810**
- max-product, **614, 713**
- max-product belief propagation, **800**
- maxent, **289**
- maximal branching, **913**
- maximal clique, **310**
- maximal information coefficient, **60**
- maximal weight bipartite matching, **659**
- maximizer of the posterior marginals, **612**
- maximum a posteriori, **4**
- maximum entropy, **39, 104, 289, 667**
- maximum entropy classifier, **252**
- maximum entropy Markov model, **685**
- maximum expected utility principle, **177**
- maximum likelihood estimate, **69**
- maximum risk, **196**
- maximum weight spanning tree, **912**
- MCAR, **270**
- MCEM, **368**
- MCMC, **52, 596, 600, 815, 837**
- MDL, **162**
- MDP, **331**
- MDS, **496**
- mean, **33**
- mean absolute deviation, **511**
- mean average precision, **303**
- mean field, **735, 756, 767, 989**
- mean field energy functional, **779**
- mean function, **291**
- mean precision, **182**
- mean reciprocal rank, **303**
- mean squared error, **205, 218**
- Mechanical Turk, **10, 995**
- median, **33**
- median model, **423**
- MEMM, **685**
- memory-based learning, **17**
- Mendelian inheritance, **317**
- Mercer kernel, **481**
- Mercer's theorem, **481, 539**
- message passing, **644, 800**
- metric, **691, 691, 803**
- metric CRF, **691**
- metric MRF, **803**

- Metropolis Hastings, **848**, 922
- Metropolis-Hastings algorithm, 869
- MFCC, 1005
- MH, **848**
- MI, **59**
- micro-averaged F1, **183**
- Microsoft, 983
- mini-batch, **264**, 571
- minimal, **282**
- minimal I-map, **324**
- minimax rule, **196**
- minimum description length, **162**
- minimum entropy prior, **621**
- minimum mean squared error, **179**
- minimum spanning tree, 897
- minorize-maximize, **369**
- misclassification loss, 176
- Misclassification rate, **547**
- misclassification rate, **22**, **205**
- missed detection, **180**
- missing, 15
- missing at random, **270**, 372, 982
- missing completely at random, **270**
- missing data, 14, 914, 974
- missing data problem, **269**
- mixed directed graphs, 931
- mixed membership model, **950**
- mixed membership stochastic block model, **973**
- mixed model, **298**
- mixing matrix, **408**
- mixing time, **857**
- mixing weights, 169, **338**
- mixture, **72**
- mixture density network, **344**
- mixture model, 164, **338**
- mixture of conjugate priors, **169**
- mixture of experts, **342**, 563, 973, 984
- mixture of factor analysers, **386**
- mixture of Gaussians, **339**
- mixture of Kalman filters, **831**
- mixture of trees, **914**
- mixture proposal, **853**
- MLE, **69**
- MLP, **563**
- MM, **369**
- MMSE, **179**
- MNIST, 7, 341
- Mobious numbers, **784**
- mode, **4**
- model based clustering, **11**
- model selection, 10, **24**, **156**
- model selection consistent, **439**
- model-based approach, xxvii
- model-based clustering, **879**
- moderated output, **260**
- modularity, xxviii
- MoE, 342
- moment matching, 176, **287**, **653**, 658, **677**
- moment parameters, **115**
- moment projection, **733**
- momentum, **248**
- monks, 974
- Monte Carlo, **52**, 151, 192, 258, 815
- Monte Carlo EM, **368**
- Monte Carlo integration, 53
- Monte Carlo localization, **828**
- moralization, **663**, **715**
- motes, 218
- motif, **36**
- mPCA, **948**
- MPE, **614**
- MPM, **612**
- MRF, **661**
- MSE, **218**
- multi label classification, 970
- multi net, **627**
- multi-armed bandit, **184**
- multi-class logistic regression, **104**
- multi-clust, **904**
- multi-grid techniques, 775
- multi-information, **415**
- multi-label classification, 3, 405
- multi-layer perceptron, **563**, 999
- multi-level model, **171**
- multi-level modeling, 844
- multi-stage, 186
- multi-target tracking, **659**
- multi-task feature selection, **297**
- multi-task learning, **172**, 231, **296**, 449, 757
- multiclass classification, 3
- multidimensional scaling, 496
- multinomial, **35**
- multinomial coefficient, **35**
- multinomial logistic regression, **104**, **252**
- multinomial PCA, **948**, 951
- multinomial probit, **295**
- multinomial regression LDA, **968**
- multinomial resampling, **826**
- multinoulli distribution, **35**
- multiple hypothesis testing, **184**
- multiple hypothesis tracking, **656**
- multiple imputation, **115**
- multiple kernel learning, **524**, 543
- multiple LDA, **276**
- multiple output model, **3**
- multiple random restarts, **348**, 921
- multiple restarts, 620
- multivariate adaptive regression splines, **553**
- multivariate Bernoulli naive Bayes, **82**
- multivariate delta method, **763**
- multivariate Gamma function, 133
- multivariate gamma function, **126**
- multivariate Gaussian, **46**, **97**, 339
- multivariate normal, **46**, **97**
- multivariate probit, **295**
- multivariate Student t, **46**
- mutual information, 46, **59**, 87, 547, 912
- mutual inhibition, **564**
- mutually independent, 62
- MVN, **46**, **97**
- N-best list, **616**
- n-gram, 568
- n-gram models, **591**
- Nadaraya-Watson, **511**
- naive Bayes classifier, **82**, 88, 311
- naive Bayes classifiers, 32
- named entity extraction, **688**
- NaN, **14**
- nats, **56**
- natural exponential family, **282**
- natural gradient, 411
- natural parameters, **115**, **282**
- NDCG, **304**
- nearest centroids classifier, **102**
- nearest medoid classification, **491**
- nearest neighbor, **16**
- nearest neighbor clustering, **897**
- nearest neighbor data association, **658**
- nearest shrunken centroids, **109**

- negative binomial, 624
- negative binomial distribution, **214**
- negative examples, 65
- negative log likelihood, **218, 349**
- negative transfer, **297**
- negentropy, **415**
- neighbors, **309**
- neocognitron, **566**
- nested plate, **321**
- Nesterov's method, **446**
- Netflix, 15, 580, 979, 981, 987, 993
- NETtalk, **569**
- neural network, 302, 969
- neural networks, 344, 535
- neutral process, **882**
- Newton's algorithm, **249, 251**
- NHST, **213**
- NIW, **133**
- NIX, **136**
- NLL, **218, 349**
- NMAR, **270**
- NMF, **470, 949**
- no forgetting, **331**
- no free lunch theorem, **24, 582**
- nodes, **309**
- nodes that fire together should wire together, **929**
- noise floor, **230**
- noisy-OR, **313, 928**
- nominal, **2**
- non-descendants, **327**
- non-factorial, **466**
- non-informative, **165**
- non-negative matrix factorization, **470, 949**
- non-negative sparse coding, **470**
- non-null recurrent, **599**
- non-parametric Bayes, **879**
- non-parametric bootstrap, **192**
- non-parametric BP, **712**
- non-parametric model, **16**
- non-parametric prior, **879**
- non-serial dynamic programming, **717**
- non-smooth, **432**
- non-terminals, 689
- nonparanormal, **942**
- norm of a function, **539**
- normal, **20, 38**
- normal equation, **220**
- normal Gamma, **476**
- normal inverse chi-squared, **136**
- Normal-inverse-wishart, **133**
- normalized cut, **891**
- normalized discounted cumulative gain, **304**
- normalized mutual information, **879**
- not missing at random, **270**
- noun phrase chunking, **687**
- NP-complete, 920
- NP-hard, 726
- ν -SVM classifier, 502
- nuisance variables, **320**
- null hypothesis, **163, 213**
- null hypothesis significance testing, **213**
- number game, **65**
- numerical underflow, **86**
- object detection, **8**
- object localization, **8**
- observation, **603**
- observation model, **312, 631**
- observed data log likelihood, **348**
- observed information, **167**
- observed information matrix, **193**
- Occam factor, **255**
- Occam's razor, **67, 156, 399, 400**
- occasionally dishonest casino, **606**
- occupancy grid, **828**
- Octave, **xxviii**
- offline, **261**
- oil wild-catter, **328**
- OLS, **220**
- OMP, **428**
- one-armed bandit, **184**
- one-hot encoding, **35**
- one-of-C encoding, 252
- one-shot decision problem, **186**
- one-standard error rule, **208**
- one-step-ahead predictive density, **609**
- one-versus-one, **503**
- one-versus-the-rest, **503**
- one-vs-all, **503**
- online EM, **365**
- online gradient descent, **262**
- online learning, **75, 241, 261**
- ontological uncertainty, **973**
- ontology, **977**
- open class, 596, **688**
- Open Directory Project, 600, 689
- open universe, **676**
- optimal action, 177
- optimism of the training error, **206**
- optimization, 218
- ordered Markov property, **310, 327**
- ordinal, 295
- ordinal regression, **2, 295, 301**
- ordinal variables, 876
- ordinary least squares, **220**
- Ornstein-Uhlenbeck process, **483**
- orthodox statistics, **191**
- orthogonal least squares, **427**
- orthogonal matching pursuits, **428**
- orthogonal projection, **221**
- out-degree, **310**
- out-of-clique query, **722**
- outer approximation, **780**
- outliers, 179, **223**
- over-complete, **282, 1001**
- overcomplete, **469**
- overcounting number, **784**
- overdispersed, **859**
- overfit, **22**
- overfitting, 72
- overrelaxed EM algorithm, **369**
- p-value, 138, 163, **163, 213**
- PAC, **210**
- PageRank, 301, 596, **600, 601**
- paired t-test, **137**
- pairwise independent, 62
- pairwise Markov property, **662**
- pairwise MRF, **666**
- parallel tempering, 858, **871, 922**
- parameter, 176
- parameter expansion, 736
- parameter modularity, **918**
- parameter sharing, **107**
- parameter tying, **107, 171, 589**
- parametric bootstrap, **192**
- parametric model, **16, 19**
- parents, **309, 310**
- Pareto distribution, **43**
- part of speech, **605, 966**

- Part of speech tagging, **605**
- partial dependence plot, **586**
- partial least squares, **406**, 975
- partially directed acyclic graph, **915**
- partially labeled LDA, **969**
- partially observed Markov decision process, **331**
- partially observed MRF, **672**
- Particle filtering, **823**
- particle filtering, 267, 648, 823, 887
- partition function, **282**, **666**
- partitional clustering, **875**
- partitioned inverse formula, **116**
- partitioning, 841
- partitions of the integers, **885**
- Parzen window density estimator, **508**
- passing a flow, **724**
- path, **310**
- path diagrams, **929**
- pathologies, **211**
- pattern, **915**
- pattern completion, **669**
- pattern recognition, **2**
- pattern search, 736, 783
- PCA, **12**, **387**, 493, 947
- PCFG, **689**
- PDAG, 936
- pdf, **32**
- pedigree graph, **315**
- peeling algorithm, **715**
- Pegasos, **701**
- penalized least squares, **226**
- penalized log likelihood, **161**
- penalized splines, **537**
- penetrance model, **317**
- perception-action, **331**
- perceptron, 569
- perceptron algorithm, **266**
- perceptual aliasing, **828**
- perfect intervention, **931**
- perfect map, **664**
- period, **598**
- permanent, 942
- perplexity, 953, **953**, 992
- persistent CD, **991**
- persistent contrastive divergence, 680
- personalized recommendation, 77
- personalized spam filtering, **296**
- perturbation theory, 892
- phase, **317**
- phase transition, **671**, 857
- phenotypes, 317
- phone, 624
- phonemes, 1005
- phylogenetic HMM, **317**
- phylogenetic tree, 925
- piecewise polynomial, **537**
- pilot runs, **851**
- pipeline, **687**
- Pitman-Koopman-Darmois theorem, **286**
- Pitman-Yor process, **885**
- Plackett-Luce, **302**
- plates, **321**
- platykurtic, **413**
- PLS, **406**
- PLSI, **949**
- plug-in, 147
- plug-in approximation, **72**
- plutocracies, 43
- pmf, **28**
- PMTK, **xxviii**
- point estimate, **149**, **150**
- pointwise approach, **301**
- pointwise marginal credibility intervals, **114**
- pointwise mutual information, **59**
- Poisson, **37**
- poisson regression, **292**
- polar, 51
- policy, **177**
- Polya urn, **89**, **884**
- Polyak-Ruppert averaging, **263**
- polynomial kernel, **481**
- polynomial regression, **20**
- polynomial time approximation schemes, 728
- polysemy, **951**
- polytree, **310**
- POMDP, **331**
- pooled, 171
- pooled empirical variance, **108**
- population minimizer, **556**
- positive definite, 125, **222**
- positive definite kernel, **481**
- positive examples, 65
- posterior expected loss, **177**
- posterior mean, 179
- posterior median, 179
- posterior mode, 178
- posterior predictive density, 608
- posterior predictive distribution, **66**, 71, 234
- potential function, **665**
- Potts model, **671**, 856
- power law, **43**
- power method, **603**
- PPCA, **381**, **387**
- precision, **38**, **182**
- precision at k, **303**, 702
- precision matrix, **46**, 100
- precision recall curve, **182**
- predict-update cycle, **609**
- predict-update-project, **653**
- predictive, **2**
- preferences, **185**
- preposterior risk, **195**
- prevalence, **183**
- Prim's algorithm, 912
- primal variables, **492**, 499
- principal component, 388
- principal components, 1000
- principal components analysis, **12**, **387**
- principal components regression, **230**
- principle of insufficient reason, **58**
- probabilistic decision tree, 551
- probabilistic expert system, **313**
- probabilistic inference, **319**
- probabilistic latent semantic indexing, **949**
- probabilistic matrix factorization, 337, **980**
- probabilistic PCA, **387**
- probabilistic principal components analysis, **381**
- probabilistic relational modeling, **675**, **976**
- probability density function, **32**
- probability mass function, **28**
- probability of the evidence, **319**, **609**, **717**
- probability product kernel, **485**
- probability simplex, **47**, 79
- probability theory, xxvii, 1
- probably approximately correct, **210**
- probe, **583**
- probit, **260**, 655
- probit regression, **293**, 362, 380, 795, 864
- product of experts, **983**
- product rule, **29**

- production rules, 689
- profile HMM, **606**
- profile log likelihood, **401**
- projected gradient descent, 444, **445**
- projection, 262
- projection pursuit, **415**
- Prolog, 676
- proposal distribution, **817**, **848**, 869
- propose, 848
- prosecutor's fallacy, **61**
- Protein sequence alignment, **606**
- protein-protein interaction networks, 970
- prototype, **341**
- proximal operator, **443**
- pruning, **549**
- pseudo counts, **75**
- pseudo likelihood, **678**
- pseudo marginals, **780**
- pseudo random number generator, **816**
- pseudo-likelihood, 943
- pure, **546**, **548**
- purity, **877**
- pushing sums inside products, **715**
- pyramid match kernel, **484**

- QALY, **186**
- QMR, **313**
- QP, **431**
- qq-plot, **260**
- QR decomposition, **228**
- quadratic discriminant analysis, **102**
- quadratic loss, **179**
- quadratic program, **431**, 498, 499
- quantile, **33**
- quantize, **59**
- quartiles, **33**
- Quasi-Newton, **251**
- query logs, 301
- query variables, **320**
- quick medical reference, **313**

- radar, 658
- radial basis function, **480**
- Rand index, **878**
- random accelerations model, **633**
- random effects, **298**
- random effects mixture of experts, **969**
- random forests, **551**, 554
- random probability measure, **880**
- random utility model, **294**
- random walk Metropolis algorithm, **848**
- random walk on the integers, 599
- random walk proposal, 869
- Rank correlation, **304**
- rank one update, **118**
- ranking, **87**, **601**, 702
- RankNet, **302**
- Rao-Blackwell, 841
- Rao-Blackwellisation, **841**
- Rao-Blackwellised particle filtering, **831**
- Rao-Blackwellized particle filtering, 659
- rare event, 182, **820**
- rate, **355**
- rational behavior, **177**
- RBF, **480**
- RBF kernel, 517
- RBF network, **486**
- RBM, **983**, 996
- RBPF, **831**
- real AdaBoost, **559**
- recall, **181**, **182**
- receiver operating characteristic, **181**
- receptive fields, **565**
- recognition weights, **410**, **986**
- recombination model, **317**
- reconstruction error, **354**, **387**
- recurrent, **599**
- recurrent neural network, **568**, **669**
- recurrent neural networks, 591
- recursive, **929**
- recursive least squares, 265, **636**
- reflecting pair, **553**
- regime switching, **660**
- regime switching Markov model, **626**
- regression, **2**
- regression spline, **537**
- regret, **262**
- regular, **598**
- regularization, **227**
- regularization path, **436**, 442, 562
- regularized discriminant analysis, **107**
- regularized estimation, **130**
- regularized particle filter, **827**
- regularized risk minimization, **206**
- reinforcement learning, **2**, 186
- reject action, **178**
- rejection sampling, **817**
- rejuvenation, **825**
- relation, **975**
- relational probabilistic models, 676
- relational topic model, **974**
- relative entropy, **57**
- relative importance of predictor variables, **586**
- relative risk, **531**
- relevance network, **908**
- relevance vector machine, 463, **488**
- Rephil, **928**
- replicated softmax model, **992**
- representer theorem, **539**
- reproducing kernel Hilbert space, **539**
- reproducing property, **539**
- rerank, 616
- resample-move, **827**
- residual, **641**
- residual analysis, **260**
- residual belief propagation, **774**
- residual error, **19**
- residual resampling, **826**
- residual sum of squares, **218**
- response variable, **2**
- responsibility, **340**, **351**
- restricted Boltzmann machine, **983**
- reverse KL, 733
- reversible jump MCMC, 370, 399, **855**
- reward, **2**
- Ricatti equations, **642**
- rich get richer, **755**, **885**
- ridge regression, 203, **226**
- right censored, **379**
- risk, **195**, 261
- risk averse, 4, **178**
- RJMCMC, **855**
- RKHS, **539**
- RLS, **636**
- Robbins-Monro, **263**, 366, 701
- robust, 179
- robust priors, **168**
- robustness, **223**
- ROC, **181**
- rocking, **261**

- root, **309**, 895
- root mean square error, 979
- Rosenblatt, 266
- rotamers, 690
- RTS smoother, **644**
- rule of iterated expectation, **141**
- rule of total probability, **29**
- rules, **550**
- RUM, **294**
- running intersection property, **722**
- RVM, **488**, 505
- saddle point approximation, **255**
- sample impoverishment, **826**
- sample standard deviation, **136**
- samples, 52
- sampling distribution, 191, **191**
- sampling importance resampling, **823**
- sampling period, **633**
- satisfying assignment, 727
- saturated model, **428**
- SBL, **463**
- scalar product, **19**
- scale invariant prior, **168**
- scale of evidence, 163
- scatter plot, **6**
- SCFGs, 624
- schedule, **263**
- Schur complement, **116**
- scientific method, 71
- scope, **328**
- score function, **167**, **193**
- score matching, 1001
- score vector, **485**
- scores, **382**
- scree plot, **400**
- screening, **87**
- search engine optimization, **603**
- second order, **249**
- second order Markov chain, **312**
- second-order Markov model, 591
- self loops, **309**
- semantic hashing, **1003**
- semantic network, 977
- semantic role labeling, 576
- semi-conjugate, **132**
- semi-continuous HMM, **630**
- semi-Markov model, **622**
- semi-metric, **691**
- semi-parametric model, **298**, **524**
- semi-supervised, 405
- semi-supervised embedding, **576**
- semi-supervised learning, **268**, 270
- sensible PCA, **387**
- sensitivity, **29**, **181**
- sensitivity analysis, **166**
- sensor fusion, **122**
- sentiment analysis, **967**
- separating set, **723**
- separation oracle, **699**
- sequence logo, **36**
- sequential, 186
- sequential minimal optimization, **499**
- sequential TRBP, **801**
- SGD, **262**
- Shafer-Shenoy, **722**
- shallow parsing, **687**
- shared, **103**
- Sherman-Morrison-Woodbury formula, **118**
- shooting, **441**, 940
- shrinkage, **122**, **174**, **230**, **557**
- shrinkage estimation, **130**
- shrinkage factor, **437**
- side chains, 690
- side information, **982**
- SIFT, 484
- sifting property, **39**
- sigma points, **650**, 651
- sigmoid, **21**, 105
- sigmoid belief net, **313**, **996**
- sigmoid belief nets, 763
- sigmoid kernel, **482**
- signal detection theory, 106
- signal processing, 421
- signal-to-noise ratio, **122**
- signal-to-symbol, **1007**
- similar, 66, 875
- similarity-based clustering, **875**
- simple cells, 413
- Simple linear regression, **241**
- simplex factor model, **949**
- Simpon's paradox, **933**
- Simulated annealing, **869**
- simulated annealing, 262, 348, **853**, 921
- simulation based, **823**
- simultaneous localization and mapping, **635**
- single best replacement, **427**
- single link clustering, **897**
- single site updating, **847**
- singular value decomposition, **392**
- singular values, **392**
- SIR, 823
- size principle, **67**
- skewness, **413**
- skip arcs, **568**
- skip-chain CRF, **688**
- slack re-scaling, **696**
- slack variables, **498**
- SLAM, **635**, 834
- slaves, 810
- slice sampling, **865**
- sliding window detector, **8**
- slippage, 635
- slot machine, 184
- small N , large D , **421**
- SmartASS, 4
- SML, **680**
- SMO, **499**
- Smoothing, **607**
- smoothing kernel, 507, **507**
- Smoothing splines, **536**
- social networks, 970
- soft clustering, **340**, **973**
- soft margin constraints, **501**
- soft thresholding, **434**, **435**
- soft weight sharing, **575**
- softmax, **104**, 283
- source coding, **56**
- SpAM, **553**
- spam, 5
- spanning tree polytope, **786**
- SpaRSA, 445
- sparse, **15**, **421**, 621, 945, 979
- sparse Bayesian learning, **463**
- sparse boosting, **562**
- sparse coding, **469**
- sparse data problem, 77
- sparse kernel machine, 421
- sparse matrix factorization, **469**, **470**
- sparse PCA, **469**

- sparse representation, 421
- sparse vector machine, **488**
- sparsity, 41
- sparsity-promoting prior, 297
- spectral, **445**
- spectral clustering, **891**
- spectral graph theory, **891**
- speech recognition, 590, 1005
- sphereing, **142**
- spherical, **46**
- spike and slab, **424**
- spin, 668
- spline, 298
- split merge, **621**
- split variable, **224**
- square root filter, **642**
- squared error, **179**
- squared exponential kernel, **480, 517**
- squared loss, 176
- squashing function, **21**
- SSM, **631**
- SSVMs, **693**
- stability selection, **439**
- stable, **936**
- stacked denoising auto-encoder, 1001
- stacking, **580**
- standard deviation, **34**
- standard error, **56**
- standard error of the mean, **137, 208**
- standard errors, **194**
- standard model, **995**
- standard normal, **38**
- standard overcomplete representation, **776**
- standardized, 352
- Standardizing, **142**
- state, 176
- state estimation, **313**
- state space, **28**
- state space model, **631**
- state transition diagram, **590, 606**
- state transition matrix, **308**
- stationary, **589, 631**
- stationary distribution, **596, 597**
- statistical learning theory, **209**
- statistical relational AI, **675**
- statistical relational learning, **976**
- statistically significant, **213**
- steepest descent, **247, 264**
- Stein's paradox, **199**
- stemming, **81**
- step size, **247**
- stepping out, **866**
- stepwise EM, **365**
- stick-breaking construction, **883**
- sticky, **850**
- stochastic algorithm, 869
- stochastic approximation, 368
- stochastic approximation EM, **368**
- stochastic automaton, **590**
- stochastic block model, **972**
- stochastic context free grammars, 624
- stochastic EM, **368**
- stochastic gradient boosting, **584**
- stochastic gradient descent, **262, 570, 868, 981, 987**
- stochastic matrix, **307, 589**
- stochastic maximum likelihood, **680, 990**
- stochastic optimization, **262**
- stochastic process, 953
- stochastic processes, **589**
- stochastic search, 429
- stochastic volatility, **831**
- stop words, **81, 480, 952**
- stopping rule, **214**
- stratified CV, **206**
- stratified sampling, **826**
- streaming data, **261**
- StreetView, 8
- strict, 197
- strictly convex, **222**
- string kernel, **483**
- strong local optimum, **804**
- strong sampling assumption, 67
- structural EM, **925**
- structural equation model, **929**
- structural equation models, 674
- structural error, **230**
- structural risk minimization, **206**
- structural signatures, **926**
- structural support vector machines, **693**
- structural time series, **637**
- structural zeros, **672**
- structure learning, **621, 681**
- structured mean field, **740**
- structured output, **684**
- structured perceptron algorithm, **700**
- structured-output classification problems, 266
- Student *t*, 359
- Student *t* distribution, **39**
- sub-Gaussian, **413**
- subderivative, **432**
- subdifferential, **432**
- subgradient, 432, **432**
- subgraph, **310**
- subjective, **67**
- subjective probability, 310
- submodular, **802**
- subsampling, **566**
- subspace method, **647**
- sufficiency principle, **214**
- sufficient statistics, **74, 79, 281, 282, 348**
- suffix trees, 483
- sum of squared errors, **218**
- sum of squares, **220**
- sum rule, **29**
- sum-product, **614, 709**
- sum-product algorithm, **707**
- super efficient, **820**
- super-Gaussian, **413**
- supermodular, **802**
- supervised LDA, **967**
- supervised learning, 2
- supervised PCA, **405**
- support, **426**
- support vector machine, **488, 496, 569**
- support vector machines, 211
- support vectors, **496, 498, 499**
- surrogate loss, 304
- surrogate loss function, **211**
- surrogate splits, **550**
- survival of the fittest, **825**
- suspicious coincidence, 164
- suspicious coincidences, **67**
- SVD, 107, 228, **392, 980**
- SVM, 211, **488, 496**
- SVMstruct, **698, 700**
- Swendsen Wang, **866**
- switching linear dynamical system, **655, 831**
- switching state space model, **655**
- symbol grounding, **1007**
- symmetric, 849

- synchronous updates, **773**
- syntactic sugar, **321**
- synthesis view, **387**
- systematic resampling, **826**
- systems biology, **13**
- systems identification, **646**
- systolic array, **710**
- t statistic, **137**
- t-test, **137**
- tabula rasa, **165**
- tail area probabilities, **33**
- tail area probability, **213**
- TAN, **312**, **914**
- TASA, **951**
- Taylor series, **255**
- Taylor series expansion, **648**
- Taylor's theorem, **248**
- temperature, **104**
- template, **676**
- template matching, **543**
- tensor product, **553**
- tensor product basis, **538**
- terminals, **689**
- test statistic, **163**, **213**
- TF-IDF, **480**
- thin junction tree filter, **635**
- thin junction trees, **944**
- thin plate spline, **538**
- thin SVD, **392**
- thinning, **862**
- Thompson sampling, **185**
- tied, **103**, **565**, **997**
- tied-mixture HMM, **630**
- Tikhonov regularization, **124**
- time reversible, **599**
- time-invariant, **589**
- time-series forecasting, **637**, **673**
- Tobit model, **379**
- Toeplitz, **627**
- tokens, **945**
- topic, **946**, **951**
- topic model, **757**
- topological ordering, **310**, **310**
- total ordering, **310**
- trace, **99**
- trace plot, **859**
- trace trick, **99**
- traceback, **614**, **717**
- tracking, **823**
- tracking by detection, **830**
- tractable substructure, **739**
- trail, **310**
- training set, **2**
- trans-dimensional MCMC, **855**
- transfer function, **563**, **570**
- transfer learning, **296**
- transient, **599**
- transition matrix, **589**, **590**
- transition model, **312**, **631**
- translation invariance, **565**, **1004**
- translation invariant, **472**
- translation invariant prior, **167**
- TRBP, **787**
- TRBP-S, **801**
- tree, **310**
- tree EP, **793**
- tree reparameterization, **774**
- tree reweighted belief propagation, **786**
- tree-augmented naive Bayes classifier, **312**
- treewidth, **320**, **719**, **800**
- trellis, **614**
- trellis diagram, **612**
- tri-cube kernel, **508**
- triangle inequality, **352**, **875**
- triangulated, **722**
- tridiagonal, **114**
- trigram model, **591**
- true positive rate, **181**
- TrueSkill, **654**, **793**
- truncated Gaussian, **362**
- truncated Gaussian potential, **691**
- truncated Newton, **250**
- truncated SVD, **393**
- TRW, **787**
- TRW-S, **801**
- tube, **497**
- tuples, **975**
- turbo codes, **768**
- two-filter smoothing, **646**
- two-slice marginal, **611**
- type I, **213**
- type I error rate, **181**
- type II maximum likelihood, **157**
- type-II maximum likelihood, **173**
- U-shaped curve, **23**
- UCB, **185**
- UGM, **661**
- UKF, **650**
- unbiased, **200**
- uncertainty, **27**
- unclamped phase, **988**
- unclamped term, **677**
- unconditionally independent, **30**
- underfits, **23**
- undirected, **309**
- undirected graphical model, **661**
- undirected local Markov property, **662**
- unfaithful, **663**
- unidentifiable, **200**, **278**, **841**
- Unified Medical Language System, **977**
- uniform distribution, **32**
- unigram statistics, **591**
- unigrams, **953**
- uninformative, **165**
- union bound, **209**
- unit information prior, **236**
- universal approximator, **564**
- unk, **81**, **596**
- unknown, **15**, **81**
- unrolled, **321**
- unscented Kalman filter, **523**, **650**
- unscented particle filter, **828**
- unscented transform, **650**
- unstable, **550**
- unsupervised learning, **2**, **9**, **337**
- up-down, **998**
- user rating profile, **949**
- utilities, **294**
- utility function, **177**
- utility nodes, **328**
- v-structure, **324**, **326**
- validation set, **23**
- value nodes, **328**
- value of perfect information, **331**
- vanishing gradient, **999**
- Vapnik-Chervonenkis, **210**
- VAR, **673**

- variable duration HMM, **622**
- variable elimination, 318, 331, **715**
- variance, **33**
- variance stabilizing transform, **175**
- variation of information, **879**
- variational Bayes, **742**
- variational Bayes EM, 620, **750**, 923
- variational EM, **368**
- variational free energy, **733**
- variational inference, 281, 318, **731**
- variational message passing, **756**
- varimax, **385**, 410
- VB, **742**
- VBEM, **750**
- VC, **210**
- VC dimension, 206
- vector auto-regressive, **673**
- vector quantization, **354**
- version space, **67**
- vertices, **309**
- VIBES, **756**
- views, **904**
- visible, 349
- visible nodes, **313**
- visible variables, **319**
- visual words, **1007**
- visualizing, **12**
- Viterbi, **612**, 701
- Viterbi decoding, **608**
- Viterbi training, **620**
- VMP, **756**
- Voronoi tessellation, **18**
- VQ, **354**

- Wald, 448
- Wald interval, **212**
- warm starting, **442**
- WARP, **304**
- Watson, 4
- wavelet, 469
- wavelet transforms, 413
- weak conditionality, **215**
- weak learner, **554**
- weak marginalization, **658**
- web crawling, **600**
- web spam, **603**
- weight decay, **226**, **572**, 987
- weight function, **533**
- weight vector, **19**
- weighted approximate-rank pairwise, **304**
- weighted average, 71
- weighted least squares, 358
- weighted least squares problem, **251**
- Whitening, **142**
- whitening, 410
- Widrow-Hoff rule, **265**
- Wishart, **125**
- working response, **250**
- World Health Organization, 60
- wrapper method, **427**

- Xbox, 654, 795
- xor, **486**

- Zellner's g-prior, 405
- zero avoiding, **733**
- zero count problem, **77**
- zero forcing, **733**
- zero temperature limit, **800**
- zig-zag, **248**