Bibliography

Abend, K., T. J. Harley, and L. N. Kanal (1965). Classification of Binary Random Patterns. *IEEE Transactions on Information Theory 11(4)*, 538–544.

Ackley, D., G. Hinton, and T. Sejnowski (1985). A learning algorithm for boltzmann machines. *Cognitive Science 9*, 147–169.

Adams, R. P., H. Wallach, and Z. Ghahramani (2010). Learning the structure of deep sparse graphical models. In *AI/Statistics*.

Aggarwal, D. and S. Merugu (2007). Predictive discrete latent factor models for large scale dyadic data. In *Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining*.

Ahmed, A. and E. Xing (2007). On tight approximate inference of the logistic-normal topic admixture model. In *AI/Statistics*.

Ahn, J.-H. and J.-H. Oh (2003). A Constrained EM Algorithm for Principal Component Analysis. *Neural Computation 15*, 57–65.

Ahn, S., A. Korattikara, and M. Welling (2012). Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring. In *Intl. Conf. on Machine Learning*.

Airoldi, E., D. Blei, S. Fienberg, and E. Xing (2008). Mixed-membership stochastic blockmodels. *J. of Machine Learning Research 9*, 1981–2014.

Aitchison, J. (1982). The statistical analysis of compositional data. *J. of Royal Stat. Soc. Series B 44*(2), 139–177.

Aji, S. M. and R. J. McEliece (2000, March). The generalized distributive law. *IEEE Trans. Info. Theory 46*(2), 325–343.

Alag, S. and A. Agogino (1996). Inference using message propogation and topology transformation in vector Gaussian continuous networks. In *UAI*.

Albers, C., M. Leisink, and H. Kappen (2006). The Cluster Variation Method for Efficient Linkage Analysis on Extended Pedigrees. *BMC Bioinformatics 7*.

Albert, J. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *J. of the Am. Stat. Assoc. 88*(422), 669–679.

Allwein, E., R. Schapire, and Y. Singer (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. *J. of Machine Learning Research*, 113–141.

Aloise, D., A. Deshpande, P. Hansen, and P. Popat (2009). NP-hardness of Euclidean sum-of-squares clustering. *Machine Learning 75*, 245–249.

Alpaydin, E. (2004). *Introduction to machine learning*. MIT Press.

Altun, Y., T. Hofmann, and I. Tsochantaridis (2006). Large Margin Methods for Structured and Interdependent Output Variables. In G. Bakir, T. Hofmann, B. Scholkopf, A. Smola, B. Taskar, and S. Vishwanathan (Eds.), *Machine Learning with Structured Outputs*. MIT Press.

Amir, E. (2010). Approximation Algorithms for Treewidth. *Algorithmica 56*(4), 448.

Amir, E. and S. McIlraith (2005). Partition-based logical reasoning for first-order and propositional theories. *Artificial Intelligence 162*(1), 49–88.

Ando, R. and T. Zhang (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *J. of Machine Learning Research 6*, 1817–1853.

Andrews, D. and C. Mallows (1974). Scale mixtures of Normal distributions. *J. of Royal Stat. Soc. Series B 36*, 99–102.

Andrieu, C., N. de Freitas, and A. Doucet (2000). Sequential Bayesian estimation and model selection for dynamic kernel machines. Technical report, Cambridge Univ.

Andrieu, C., N. de Freitas, and A. Doucet (2001). Robust Full Bayesian Learning for Radial Basis Networks. *Neural Computation 13*(10), 2359–2407.

Andrieu, C., N. de Freitas, A. Doucet, and M. Jordan (2003). An introduction to MCMC for machine learning. *Machine Learning 50*, 5–43.

Andrieu, C., A. Doucet, and V. Tadic (2005). Online EM for parameter estimation in nonlinear-non Gaussian state-space models. In *Proc. IEEE CDC*.

Andrieu, C. and J. Thoms (2008). A tutorial on adaptive MCMC. *Statistical Computing 18*, 343–373.

Aoki, M. (1987). *State space modeling of time series*. Springer.

Archambeau, C. and F. Bach (2008). Sparse probabilistic projections. In *NIPS*.

Argyriou, A., T. Evgeniou, and M. Pontil (2008). Convex multi-task feature learning. *Machine Learning 73*(3), 243–272.

Armagan, A., D. Dunson, and J. Lee (2011). Generalized double pareto shrinkage. Technical report, Duke.

Armstrong, H. (2005). *Bayesian estimation of decomposable Gaussian graphical models*. Ph.D. thesis, UNSW.

Armstrong, H., C. Carter, K. Wong, and R. Kohn (2008). Bayesian Covariance Matrix Estimation using a Mixture of Decomposable Graphical Models. *Statistics and Computing*, 1573–1375.

Arnborg, S., D. G. Corneil, and A. Proskurowski (1987). Complexity of finding embeddings in a k-tree. *SIAM J. on Algebraic and Discrete Methods 8*, 277–284.

Arora, S. and B. Barak (2009). *Complexity Theory: A Modern Approach*. Cambridge.

Arthur, D. and S. Vassilvitskii (2007). k-means++: the advantages of careful seeding. In *Proc. 18th ACM-SIAM symp. on Discrete algorithms*, pp. 1027âĂŞ1035.

Arulampalam, M., S. Maskell, N. Gordon, and T. Clapp (2002, February). A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking. *IEEE Trans. on Signal Processing 50*(2), 174–189.

Asavathiratham, C. (2000). *The Influence Model: A Tractable Representation for the Dynamics of Networked Markov Chains*. Ph.D. thesis, MIT, Dept. EECS.

Atay-Kayis, A. and H. Massam (2005). A Monte Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika 92*, 317–335.

Attenberg, J., K. Weinberger, A. Smola, A. Dasgupta, and M. Zinkevich (2009). Collaborative spam filtering with the hashing trick. In *Virus Bulletin*.

Attias, H. (1999). Independent factor analysis. *Neural Computation 11*, 803–851.

Attias, H. (2000). A variational Bayesian framework for graphical models. In *NIPS-12*.

Bach, F. (2008). Bolasso: Model Consistent Lasso Estimation through the Bootstrap. In *Intl. Conf. on Machine Learning*.

Bach, F. and M. Jordan (2001). Thin junction trees. In *NIPS*.

Bach, F. and M. Jordan (2005). A probabilistic interpretation of canonical correlation analysis. Technical Report 688, U. C. Berkeley.

Bach, F. and E. Moulines (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *NIPS*.

Bahmani, B., B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii (2012). Scalable k-Means++. In *VLDB*.

Bakker, B. and T. Heskes (2003). Task Clustering and Gating for Bayesian Multitask Learning. *J. of Machine Learning Research 4*, 83–99.

Baldi, P. and Y. Chauvin (1994). Smooth online learning algorithms for hidden Markov models. *Neural Computation 6*, 305–316.

Balding, D. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics 7*, 81–91.

Banerjee, A., S. Basu, and S. Merugu (2007). Multi-way clustering on relation graphs. In *Proc. SIAM Intl. Conf. on Data Mining (SDM)*.

Banerjee, O., L. E. Ghaoui, and A. d'Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *J. of Machine Learning Research 9*, 485–516.

Bar-Shalom, Y. and T. Fortmann (1988). *Tracking and data association*. Academic Press.

Bar-Shalom, Y. and X. Li (1993). *Estimation and Tracking: Principles, Techniques and Software*. Artech House.

Barash, Y. and N. Friedman (2002). Context-specific Bayesian clustering for gene expression data. *J. Comp. Bio. 9*, 169–191.

Barber, D. (2006). Expectation Correction for Smoothed Inference in Switching Linear Dynamical Systems. *J. of Machine Learning Research 7*, 2515–2540.

Barber, D. and C. Bishop (1998). Ensemble Learning in Bayesian Neural Networks. In C. Bishop (Ed.), *Neural Networks and Machine Learning*, pp. 215–237. Springer.

Barber, D. and S. Chiappa (2007). Unified inference for variational bayesian linear gaussian state space models. In *NIPS*.

Barbieri, M. and J. Berger (2004). Optimal predictive model selection. *Annals of Statistics 32*, 870–897.

Bartlett, P., M. Jordan, and J. McAuliffe (2006). Convexity, Classification, and Risk Bounds. *J. of the Am. Stat. Assoc. 101*(473), 138–156.

Baruniak, R. (2007). Compressive sensing. *IEEE Signal Processing Magazine*.

Barzilai, J. and J. Borwein (1988). Two point step size gradient methods. *IMA J. of Numerical Analysis 8*, 141–148.

Basu, S., T. Choudhury, B. Clarkson, and A. Pentland (2001). Learning human interactions with the influence model. Technical Report 539, MIT Media Lab. ftp://whitechapel.media.mit.edu/pub/tech-reports/TR-539-ABSTRACT.html.

Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970). A maximization technique occuring in the statistical analysis of probabalistic functions in markov chains. *The Annals of Mathematical Statistics 41*, 164–171.

Beal, M. (2003). *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, Gatsby Unit.

Beal, M. and Z. Ghahramani (2006). Variational Bayesian Learning of Directed Graphical Models with Hidden Variables. *Bayesian Analysis 1*(4).

Beal, M. J., Z. Ghahramani, and C. E. Rasmussen (2002). The infinite hidden Markov model. In *NIPS-14*.

Beck, A. and M. Teboulle (2009). A fast iterative shrinkage-thresholding algorothm for linear inverse problems. *SIAM J. on Imaging Sciences 2*(1), 183–202.

Beinlich, I., H. Suermondt, R. Chavez, and G. Cooper (1989). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proc. of the Second European Conf. on AI in Medicine*, pp. 247–256.

Bekkerman, R., M. Bilenko, and J. Langford (Eds.) (2011). *Scaling Up Machine Learning*. Cambridge.

Bell, A. J. and T. J. Sejnowski (1995). An information maximisation approach to blind separation and blind deconvolution. *Neural Computation 7*(6), 1129–1159.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning 2*(1), 1–127.

Bengio, Y. and S. Bengio (2000). Modeling high-dimensional discrete data with multi-layer neural networks. In *NIPS*.

Bengio, Y., O. Delalleau, N. Roux, J. Paiement, P. Vincent, and M. Ouimet (2004). Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation 16*, 2197–2219.

Bengio, Y. and P. Frasconi (1995). Diffusion of context and credit information in markovian models. *J. of AI Research 3*, 249–270.

Bengio, Y. and P. Frasconi (1996). Input/output HMMs for sequence processing. *IEEE Trans. on Neural Networks 7*(5), 1231–1249.

Bengio, Y., P. Lamblin, D. Popovici, and H. Larochelle (2007). Greedy layer-wise training of deep networks. In *NIPS*.

Berchtold, A. (1999). The double chain markov model. *Comm. Stat. Theor. Methods 28*, 2569–2589.

Berger, J. (1985). Bayesian salesmanship. In P. K. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques with Applications: Essays in Honor of Bruno deFinetti*. North-Holland.

Berger, J. and R. Wolpert (1988). *The Likelihood Principle*. The Institute of Mathematical Statistics. 2nd edition.

Berkhin, P. (2006). A survey of clustering datamining techniques. In J. Kogan, C. Nicholas, and M. Teboulle (Eds.), *Grouping Multidimensional Data: Recent Advances in Clustering*, pp. 25–71. Springer.

Bernardo, J. and A. Smith (1994). *Bayesian Theory*. John Wiley.

Berrou, C., A. Glavieux, and P. Thitimajashima (1993). Near Shannon limit error-correcting coding and decoding: Turbo codes. *Proc. IEEE Intl. Comm. Conf.*.

Berry, D. and Y. Hochberg (1999). Bayesian perspectives on multiple comparisons. *J. Statist. Planning and Inference 82*, 215–227.

Bertele, U. and F. Brioschi (1972). *Nonserial Dynamic Programming*. Academic Press.

Bertsekas, D. (1997). *Parallel and Distribution Computation: Numerical Methods*. Athena Scientific.

Bertsekas, D. (1999). *Nonlinear Programming* (Second ed.). Athena Scientific.

Bertsekas, D. and J. Tsitsiklis (2008). *Introduction to Probability*. Athena Scientific. 2nd Edition.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician 24*, 179–196.

Bhatnagar, N., C. Bogdanov, and E. Mossel (2010). The computational complexity of estimating convergence time. Technical report, .

Bhattacharya, A. and D. B. Dunson (2011). Simplex factor models for multivariate unordered categorical data. *J. of the Am. Stat. Assoc.*. To appear.

Bickel, P. and E. Levina (2004). Some theory for Fisher's linear discriminant function, "Naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli 10*, 989–1010.

Bickson, D. (2009). *Gaussian Belief Propagation: Theory and Application*. Ph.D. thesis, Hebrew University of Jerusalem.

Bilmes, J. (2000). Dynamic Bayesian multinets. In *UAI*.

Bilmes, J. A. (2001). Graphical models and automatic speech recognition. Technical Report UWEETR-2001-0005, Univ. Washington, Dept. of Elec. Eng.

Binder, J., D. Koller, S. J. Russell, and K. Kanazawa (1997). Adaptive probabilistic networks with hidden variables. *Machine Learning 29*, 213–244.

Binder, J., K. Murphy, and S. Russell (1997). Space-efficient inference in dynamic probabilistic networks. In *Intl. Joint Conf. on AI*.

Birnbaum, A. (1962). On the foundations of statistical infernece. *J. of the Am. Stat. Assoc. 57*, 269–326.

Bishop, C. (1999). Bayesian PCA. In *NIPS*.

Bishop, C. (2006a). *Pattern recognition and machine learning*. Springer.

Bishop, C. (2006b). *Pattern recognition and machine learning*. Springer.

Bishop, C. and G. James (1993). Analysis of multiphase flows using dual-energy densitometry and neural networks. *Nuclear Instruments and Methods in Physics Research A327*, 580–593.

Bishop, C. and M. Svensén (2003). Bayesian hierarchical mixtures of experts. In *UAI*.

Bishop, C. and M. Tipping (2000). Variational relevance vector machines. In *UAI*.

Bishop, C. M. (1994). Mixture density networks. Technical Report NCRG 4288, Neural Computing Research Group, Department of Computer Science, Aston University.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press.

Bishop, Y., S. Fienberg, and P. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

Bistarelli, S., U. Montanari, and F. Rossi (1997). Semiring-based constraint satisfaction and optimization. *J. of the ACM 44*(2), 201–236.

Blake, A., P. Kohli, and C. Rother (Eds.) (2011). *Advances in Markov Random Fields for Vision and Image Processing*. MIT Press.

Blei, D. and J. Lafferty (2006a). Correlated topic models. In *NIPS*.

Blei, D. and J. Lafferty (2006b). Dynamic topic models. In *Intl. Conf. on Machine Learning*, pp. 113–120.

Blei, D. and J. Lafferty (2007). A Correlated Topic Model of "Science". *Annals of Applied Stat. 1*(1), 17–35.

Blei, D. and J. McAuliffe (2010, March). Supervised topic models. Technical report, Princeton.

Blei, D., A. Ng, and M. Jordan (2003). Latent dirichlet allocation. *J. of Machine Learning Research 3*, 993–1022.

Blumensath, T. and M. Davies (2007). On the difference between Orthogonal Matching Pursuit and Orthogonal Least Squares. Technical report, U. Edinburgh.

Bo, L., C. Sminchisescu, A. Kanaujia, and D. Metaxas (2008). Fast Algorithms for Large Scale Conditional 3D Prediction. In *CVPR*.

Bohning, D. (1992). Multinomial logistic regression algorithm. *Annals of the Inst. of Statistical Math. 44*, 197–200.

Bollen, K. (1989). *Structural Equation Models with Latent Variables*. John Wiley & Sons.

Bordes, A., L. Bottou, and P. Gallinari (2009, July). Sgd-qn: Careful quasi-newton stochastic gradient descent. *J. of Machine Learning Research 10*, 1737–1754.

Bordes, A., L. Bottou, P. Gallinari, J. Chang, and S. A. Smith (2010). Erratum: SGDQN is Less Careful than Expected. *J. of Machine Learning Research 11*, 2229–2240.

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992). A training algorithm for optimal margin classifiers. In *Proc. of the Workshop on Computational Learning Theory*.

Bottcher, S. G. and C. Dethlefsen (2003). deal: A package for learning bayesian networks. *J. of Statistical Software 8*(20).

Bottolo, L. and S. Richardson (2010). Evolutionary stochastic search. *Bayesian Analysis 5*(3), 583–618.

Bottou, L. (1998). Online algorithms and stochastic approximations. In D. Saad (Ed.), *Online Learning and Neural Networks*. Cambridge.

Bottou, L. (2007). Learning with large datasets (nips tutorial).

Bottou, L., O. Chapelle, D. DeCoste, and J. Weston (Eds.) (2007). *Large Scale Kernel Machines*. MIT Press.

Bouchard, G. (2007). Efficient bounds for the softmax and applications to approximate inference in hybrid models. In *NIPS 2007 Workshop on Approximate Inference in Hybrid Models*.

Bouchard-Cote, A. and M. Jordan (2009). Optimization of structured mean field objectives. In *UAI*.

Bowman, A. and A. Azzalini (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford.

Box, G. and N. Draper (1987). *Empirical Model-Building and Response Surfaces*. Wiley.

Box, G. and G. Tiao (1973). *Bayesian inference in statistical analysis*. Addison-Wesley.

Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge.

Boyen, X. and D. Koller (1998). Tractable inference for complex stochastic processes. In *UAI*.

Boykov, Y., O. Veksler, and R. Zabih (2001). Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence 23*(11).

Brand, M. (1996). Coupled hidden Markov models for modeling interacting processes. Technical Report 405, MIT Lab for Perceptual Computing.

Brand, M. (1999). Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation 11*, 1155–1182.

Braun, M. and J. McAuliffe (2010). Variational Inference for Large-Scale Models of Discrete Choice. *J. of the Am. Stat. Assoc. 105*(489), 324–335.

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*, 123–140.

Breiman, L. (1998). Arcing classifiers. *Annals of Statistics 26*, 801–849.

Breiman, L. (2001a). Random forests. *Machine Learning 45*(1), 5–32.

Breiman, L. (2001b). Statistical modeling: the two cultures. *Statistical Science 16*(3), 199–231.

Breiman, L., J. Friedman, and R. Olshen (1984). *Classification and regression trees*. Wadsworth.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *J. of the Am. Stat. Assoc. 88*(421), 9–25.

Briers, M., A. Doucet, and S. Maskel (2010). Smoothing algorithms for state-space models. *Annals of the Institute of Statistical Mathematics 62*(1), 61–89.

Brochu, E., M. Cora, and N. de Freitas (2009, November). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Technical Report TR-2009-23, Department of Computer Science, University of British Columbia.

Brooks, S. and G. Roberts (1998). Assessing convergence of Markov Chain Monte Carlo algorithms. *Statistics and Computing 8*, 319–335.

Brown, L., T. Cai, and A. DasGupta (2001). Interval estimation for a binomial proportion. *Statistical Science 16*(2), 101–133.

Brown, M. P., R. Hughey, A. Krogh, I. S. Mian, K. Sjölander, and D. Haussler (1993). Using dirichlet mixtures priors to derive hidden Markov models for protein families. In *Intl. Conf. on Intelligent Systems for Molecular Biology*, pp. 47–55.

Brown, P., M. Vannucci, and T. Fearn (1998). Multivariate Bayesian variable selection and prediction. *J. of the Royal Statistical Society B 60*(3), 627–641.

Bruckstein, A., D. Donoho, and M. Elad (2009). From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review 51*(1), 34–81.

Bryson, A. and Y.-C. Ho (1969). *Applied optimal control: optimization, estimation, and control*. Blaisdell Publishing Company.

Buhlmann, P. and T. Hothorn (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science 22*(4), 477–505.

Buhlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methodology, Theory and Applications*. Springer.

Buhlmann, P. and B. Yu (2003). Boosting with the L2 loss: Regression and classification. *J. of the Am. Stat. Assoc. 98*(462), 324–339.

Buhlmann, P. and B. Yu (2006). Sparse boosting. *J. of Machine Learning Research 7*, 1001–1024.

Bui, H., S. Venkatesh, and G. West (2002). Policy Recognition in the Abstract Hidden Markov Model. *J. of AI Research 17*, 451–499.

Buntine, W. (2002). Variational Extensions to EM and Multinomial PCA. In *Intl. Conf. on Machine Learning*.

Buntine, W. and A. Jakulin (2004). Applying Discrete PCA in Data Analysis. In *UAI*.

Buntine, W. and A. Jakulin (2006). Discrete Component Analysis. In *Subspace, Latent Structure and Feature Selection: Statistical and Optimization Perspectives Workshop*.

Buntine, W. and A. Weigend (1991). Bayesian backpropagation. *Complex Systems 5*, 603–643.

Burges, C. J., T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender (2005). Learning to rank using gradient descent. In *Intl. Conf. on Machine Learning*, pp. 89–96.

Burkard, R., M. Dell'Amico, and S. Martello (2009). *Assignment Problems*. SIAM.

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 l1 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. 129âĂŞ136.

Cappe, O. (2010). Online Expectation Maximisation. In K. Mengersen, M. Titterington, 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*, 165âĂŞ188.

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, Pratetsky-Shapiro, Smyth, and Uthurasamy (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. SkÃűld (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. Multiscale 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.

Cowell, R. G., A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. Springer.

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 learnng 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. Yavnch (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 fmaily approximation of the Dirichlet compoind 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.), *Mathemetical and Statistica 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), 907âĂŞ917.

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*, 972âĂŞ976.

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, Februrary). 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 verus the lasso. *J. Computational and graphical statistics*.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition*. Academic Press. 2nd edition.

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 Dirchlet 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. *Statisical 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 analyzers. 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 infernece 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 Comptuation 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. *IEE Proceedings (F) 140*(2), 107–113.

Graepel, T., J. Quinonero-Candela, T. Borchert, and R. Herbrich (2010). Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in MicrosoftâĂŹs 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). *Nonparametric 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.

Gu, Q., Z. Li, and J. Han (2011). Generalized Fisher Score for Feature Selection. In *UAI*.

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 Univerity 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 density 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 Comptuation*. 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., L. Deng, D. Yu, G. Dahl, A.Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury (2012, November). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Proc. Magazine 29*(6).

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 satisïňĄed. In *UAI*.

Hjort, N., C. Holmes, P. Muller, and S. Walker (Eds.) (2010). *Bayesian Nonparametrics*. Cambridge.

Hoefling, H. (2010). A Path Algorithm for the Fused Lasso Signal Approximator. Technical report, Stanford.

Hoefling, 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), 2554âĂŞ2558.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks 4*(2), 251âĂŞ257.

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 factorizaton 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 itneraction, 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*, 73âĂŞ101.

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. Opper 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.

Jensen, D., J. Neville, and B. Gallagher (2004). Why collective inference improves relational classification. In *Proc. of the Int'l Conf. on Knowledge Discovery and Data Mining*.

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. Schwarts, 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 criterio. *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 noramal 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. Zabin (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. 1288âĂŞ1298.

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 Int'l 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 Lassos. *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 disribution. *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 examplar-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 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). Hyperspacings 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 (2009). 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 171*(5), 311–331.

Lindley, D. (1982). Scoring rules and the inevetability 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 Identificiation: 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. *Statisica 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., C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter (2012). *The BUGS Book: A practical Introduction to Bayesian Analysis*. CRC Press.

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. BÃijhlmann (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods 7*, 247–248.

Maathuis, M., M. Kalisch, and P. BÃijhlmann (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). Comparision 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. Willsky (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 Bionformatics 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, Stanford Dept. Statistics.

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.

McCullich, 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. Buhlmann (2006). High dimensional graphs and variable selection with the lasso. *The Annals of Statistics 34*, 1436–1462.

Meinshausen, N. and P. BÃijhlmann (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 statisics. Technical report, MIT Media Lab.

Minka, T. (2000a). Automatical 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. Gassiat (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 (ICASSP'97)*, 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), 1077–1087.

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.

Opper, M. (1998). A Bayesian approach to online learning. In D. Saad (Ed.), *On-line learning in neural networks*. Cambridge.

Opper, M. and C. Archambeau (2009). The variational Gaussian approximation revisited. *Neural Computation 21*(3), 786–792.

Opper, 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 nonnegative 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). Nonparametric 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.

Petris, 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 multilocus 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 Kauffman.

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.

Rattray, 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 Statisical 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 reconstrution 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), 386âĂŞ408.

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. *Pyschonomic 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.

Salojarvi, J., K. Puolamaki, 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 regresssion 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. Mueller (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), 461âĂŞ464.

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. Bohlen, N. KrÃijger, S. Sonnenburg, and G. RÃd'tsch (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. MyllymÃd'ki (2007). On Sensitivity of the MAP Bayesian Network Structure to the Equivalent Sample Size Parameter. In *UAI*, pp. 360–367.

Silander, T. and P. Myllmaki (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. Rumehart and J. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1*. 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). Minimum probability flow learning. 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*, 465âĂŞ 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*.

Spirtes, 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. Trumpf, 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). *Reinforcment 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 (2012). Efficient Bayesian Inference for Multivariate Probit Models with Sparse Inverse Correlation Matrices. *J. Comp. Graph. Statist. 21*(3), 739–757.

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 21*(3), 611–622.

Tipping, M. and A. Faul (2003). Fast marginal likelihood maximisation for sparse bayesian models. In *AI/Stats*.

Tishby, N., F. Pereira, and W. Biale (1999). The information bottleneck method. In *The 37th annual Allerton Conf. on Communication, Control, and Computing*, pp. 368âĂŞ377.

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.

van Iterson, M., H. van Haagen, and J. Goeman (2012). Resolving confusion of tongues in statistics and machine learning: A primer for biologists and bioinformaticians. *Proteomics 12*, 543âĂŞ549.

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. PietilÃď inen, 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), 260âĂŞ269.

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 $\ell - 1$-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. Fruhrwirth (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 Opper (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-lcass 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 Disrete 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 l1-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 minimze 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 counter-intuitive 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. 928âĂŞ936.

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.*