# Statistical methods for mass spectrometry-based clinical proteomics

Kakourou, A.A.

# Bibliography


Adams, N. M., D. K. Tasoulis, C. Anagnostopoulos, and D. J. Hand (2010). Temporally-adaptive linear classification for handling population drift in credit scoring. *Proceedings of COMPSTAT 2010*.

Anderson, N. L. and N. G. Anderson (2002). The human plasma protein. *Molecular & Cellular Proteomics 1*, 845–867.

Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics 19 (2)*, 185–193.

Breiman, L. (1996). Stacked regressions. *Machine learning 24*, 49–64.

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

Bühlmann, P. and T. Hothorn (2007). Boosting algorithms: regularization, prediction and model fitting. *Statistical Science 22*, 477–505.

Burzykowski, T., J. Claesen, and D. Valkenborg (2016). The analysis of peptide-centric mass-spectrometry data utilizing information about the expected isotope distribution. In *Statitcal analysis of proteomics, metabolomics and lipidomics data using mass spectromtery*. Springer.

Cox, D. (1958). Two further applications of a model for binary regression. *Biometrika 45*, 562–565.

de Noo, M., A. Deelder, B. Mertens, B. M. Ozalp, A., M. van der Werff, and R. Tollenaar (2005). Detection of colorectal cancer using maldi-tof serum protein profiling. *European Journal of Cancer 42*, 1068–1076.

Dellaportas, P., J. Forster, and I. Ntzoufras (2000). Bayesian variable selection using the gibbs sampler. *Generalized Linear Models: A Bayesian Perspective*.

Dellaportas, P., J. Forster, and I. Ntzoufras (2002). On bayesian model and variable selection using mcmc. *Statistics and Computing 12*, 27–36.

Denison, D. G. T., C. C. Holmes, B. K. Mallich, and S. A. F. M (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley series in probability and statistics.

Diamandis, E. P. (2004). Mass spectrometry as a diagnostic and a cancer biomarker discovery tool opportunities and potential limitations. *Molecular & Cellular Proteomics 3*, 367–378.

Dong, T., C. C. Liu, E. F. Petricoin, and L. L. Tang (2014). Combining markers with and without the limit of detection. *Statitics in Medicine 33(8)*, 1307–1320.

Fearn, T. (2008). Principal component discriminant analysis. *Statistical Applications in Genetics and Molecular Biology 7(2)*.

Friedman, J., T. Hastie, and R. Tibshirani (2013). *The glmnet Package for R*. Version 1.9-5.

George, E. I. G. and R. E. McCulloch (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association, 88(423)*, 881–889.

Goeman, J. (2016). *Penalized R package*. version 0.9-47.

Hand, D. (1997). Construction and assessment of classification rules. *Wiley, Chichester*.

Hand, D. (2006). Classifier technology and the illusion of progress. *Statistical Science 21*, 1–18.

Hand, D. (2008). Breast cancer diagnosis from proteomic mass spectrometry data: A comparative evaluation. *Statistical Applications in Genetics and Molecular Biology 7 (2)*.

Hastie, T., R. Tibshirani, and J. Friedman (2001). Elements of statistical learning: Data mining, inference, and prediction. *Springer Series in Statistics*.

Helsel, D. (1990). Less than obvious: Statistical treatment of data below the dtection liimit. *Enviromental Science and Technology 24(12)*, 1766–1774.

Helsel, D. R. (2012). *Statistics for Censored Enviromental Data Using MINITAB and R*. New Jersey: Wiley.

Henningsen, A. (2010). Estimating censored regression models in r using the censreg package. *University of Copenhagen*.

Hoefsloot, H. C. J., S. Smit, and A. K. Smilde (2008). A classification model for the leiden proteomics competition. *Statistical Applications in Genetics and Molecular Biology 7(2)*.

Hoerl, A. and R. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*, 55–67.

Hopke, P. K., C. Liu, and R. D. B (2001). Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the arctic. *Biometrics 57*, 22–33.

Horn, D. M., R. A. Zubarev, and F. W. McLafferty (2000). Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *Journal of the American Society for Mass Spectrometry 11*, 320–332.

Hornung, R. W. and R. L. D (2000). Estimation of average concentration in the presence of nondetectable values. *Applied Occupational Environmental Hygiene 5*, 46–51.

Hughes, J. (1982). Mixed effects models with censored data with application to hiv rna levels. *Biometrics 55(2)*, 625–629.

Hughes, J. P. (1999). Mixed effects models with censored data with application to hiv rna levels. *Biometrics 55(2)*, 625–629.

Inouye, M., J. Kettunen, P. Soininen, K. Silander, and S. Ripatti (2010). On the use of cross-validation to assess performance in multivariate prediction. *Molecular Systems Biology 6*, 441.

Inouye, M., K. Silander, E. Hamalainen, V. Salomaa, K. Harald, and P. Jousilahti (2010). An immune response network associated with blood lipid levels. *Plos Genetics 6*.

Jonathan, P., K. W. and M. McCarthy (2000). On the use of cross-validation to assess performance in multivariate prediction. *Statistics and Computing 10*, 209–229.

Kakourou, A., W. Vach, and B. Mertens (2014). Combination approaches improve predictive performance of diagnostic rules for mass-spectrometry proteomic data. *Journal of Computational Biology 21*, 898–914.

Kakourou, A., W. Vach, S. Nicolardi, Y. van der Burgt, and B. Mertens (2016). Accounting for isotopic clustering in fourier transform mass spectrometry data analysis for clinical diagnostic studies. *Statistical Applications in Genetics and Molecular Biology 15(5)*, 415–430.

Karpievitch, Y. V., A. R. Dabney, and S. R. D (2012). Normalization and missing value imputation for label-free lc-ms analysis. *BMC Bioinformatics 13(16)*.

Karpievitch, Y. V., J. Stanley, T. Taverner, J. Huang, J. N. Adkins, C. Ansong, F. Heffron, T. O. Metz, W. J. Qian, H. Yoon, R. D. Smith, and A. R. Dabney (2009). A statistical framework for protein quantitation in bottom-up ms-based proteomics. *Bioinformatics 25(16)*, 2028–2034.

Kelly, M. G., D. J. Hand, and N. M. Adams (1999). The impact of changing populations on classifier performance. *Proceedings of the Fifth ACM SIGKDD International Conference on Knowedge Discovery and Data Mining*.

Kneib, T., T. Hothorn, and G. Tutz (2009). Variable selection and model choice in geoadditive regression models. *Biometrics 65*, 626–634.

Krzanowski, W. J., P. Jonathan, W. V. McCarthy, and M. R. Thomas (1995). Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. *Applied Statistics 44(1)*, 101–115.

Laird, M. and J. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38(4)*, 963–974.

Le Cessie, S. and J. C. van Houwelingen (1992). Ridge estimators in logistic regression. *Applied Statistics 41 (1)*, 191–201.

Leblanc, M. and R. Tibshirani (1996). Combining estimates in regression and classification. *Journal of the American Statistical Association 91*, 1641–1650.

Liu, H., P. DÁndrade, S. Fulmer-Smentek, P. Lorenzi, K. Kohn, J. Weinstein, Y. Pommier, and W. Reinhold (2014). mrna and microrna expression profiles of the nci-60 integrated with drug activities. *Mol Cancer Ther 9*, 1080–1091.

Meier, L., S. van de Geer, and P. Bühlmann (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society B 70*, 83–71.

Meijer, R. and J. Goeman (2013). Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal 55(2)*, 141–155.

Mertens, B. (2003). Microarrays, pattern recognition and exploratory data analysis. *Statistics in Medicine 22*, 1879–1899.

Mertens, B. (2008). Organizing a competition on clinical mass spectrometry based proteomic diagnosis. *Statistical Applications in Genetics and Molecular Biology 7(2)*.

Mertens, B. (2016). Logistic regression modeling on mass spectrometry data in proteomics case-control discriminant studies. In *Statitcal analysis of proteomics, metabolomics and lipidomics data using mass spectromtery*. Springer.

Mertens, B., M. E. De Noo, R. A. E. M. Tollenaar, and A. M. Deelder (2006). Mass spectrometry proteomic diagnosis: Enacting the double cross-validatory paradigm. *Journal of Computational Biology 13*, 1591–1605.

Mertens, B., Y. van der Burgt, B. Velstra, W. Mesker, R. Tollenaar, and A. Deelder (2011). On the use of double cross-validationon the use of double cross-validation for the combination of proteomic mass spectral data for enhanced diagnosis and prediction. *Statistics and Probability Letters 81*, 759–766.

Morris, J. S., K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi (2005). Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics 21*, 1764–1775.

Nicolardi, S., B. J. Velstra, B. Mertens, B. Bosing, W. E. Mesker, R. A. E. M. Tollenaar, A. M. Deelder, and Y. E. M. van der Burgt (2014). Ultrahigh resolution profiles lead to more detailed serum peptidome sugnatures of pancreatic cancer. *Translational Proteomics 2*, 39–51.

Palmblad, M., J. Buijs, and P. Hakanson (2001). Automatic analysis of hydrogen/deuterium exchange mass spectra of peptides and proteins using calculations of isotopic distributions. *Journal of The American Society for Mass Spectrometry 12*, 1153–1162.

Park, K., J. Y. Yoon, S. Lee, E. Paek, H. Park, H. J. Jung, and S. W. Lee (2008). Isotopic peak intensity ratio based algorithm for determination of isotopic clusters and monoisotopic masses of polypeptides from high-resolution mass spectrometric data. *Journal of Analytical Chemistry 80*, 7294–7303.

Pencina, M., R. S. D'Agostino, R. J. D'Agostino, and R. Vasan (2010). Evaluating the added predictive ability of a new marker: from area under the roc curve to reclassification and beyond. *Statistics in Medicine 27*, 157–172.

Pepe, M., K. Kerr, G. Longton, and Z. Wang (2013). Testing for improvement in prediction model performance. *Statistics in Medicine 32*, 1467–1482.

Rendell, A. and R. Seshu (1990). Learning hard concepts through constructive induction: Framework and rationale. *Computational Intelligence 6*, 247–270.

Rockwood, A. L. and P. Haimi (2006). Efficient calculation of accurate masses of isotopic peaks. *Journal of The American Society for Mass Spectrometry 17*, 415–419.

Rodríguez-Girondo, M., P. Salo, T. Burzykowski, M. Perola, J. Houwing-Duistermaat, and B. Mertens (2016). Sequential double cross-validation for augmented prediction assessment in high-dimensional omic applications. *Working Paper in ArXiv*.

Sauve, A. C. and T. P. Speed (2004). Normalization, baseline correction and alignment of high-throughput mass spectrometry data. *Procedings Gensips*.

Scheltema, R. (2009). Simple data-reduction method for high-resolution lc-ms data in metabolomics. *Bioanalysis 1(9)*, 1551–7.

Senko, M. W., S. C. Beu, and F. W. McLafferty (1995). Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distribution. *Journal of the American Society for Mass Spectrometry 6*, 229–233.

Shavlik, J., R. Mooney, and G. Towell (1991). Symbolic and neural learning algorithms: An experimental comparison. *Machine learning 6*, 11–143.

Steyerberg, E., A. Vickers, N. Cook, T. Gerds, M. Gonen, N. Obuchowski, M. Pencina, and M. Kattan (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology 21*, 128–138.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society 3(2)*.

Strimenopoulou, F. and P. Brown (2008). Empirical bayes logistic regression. *Statistical Applications in Genetics and Molecular Biology 7(2)*.

Succop, P. A., S. Clark, and M. Chen (2004). Imputation of data values that are less than a detection limit. *Journal of Occupational and Environmental Health 1*, 436–441.

Tekwe, C. D., R. J. Carroll, and A. R. Dabney (2012). Application of survival analysis methodology to the quantitative analysis of lc-ms proteomics data. *Bioinfoamtics 28(15)*, 1998–2003.

Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.

Therneau, T. M. and G. P. M (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Tibshirani, R. (1996). Regression shrinkage and variable selection via the lasso. *Journal of the Royal Statistical Society 58*, 267–288.

Tibshirani, R. and B. Efron (2002). Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology 1(1)*.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Ecomonetrica 26 (1)*, 24–36.

Tutz, G. and H. Binder (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics 62*, 961–971.

Valkenborg, D., I. Mertens, F. Lemiere, E. Witters, and T. Burzykowski (2012). The isotopic distribution conundrum. *Mass Spectromtry Revies 31*, 96–109.

van de Wiel, M., T. Lien, W. Verlaat, W. van Wieringen, , and S. Wilting (2015). Better prediction by use of co-data: Adaptive group-regularized ridge regression. *Statistics in Medicine 35*, 368–381.

van der Burgt, Y. E. M., I. M. Taban, M. Konijnenburg, M. Biskup, M. C. Duursma, R. M. A. Heeren, A. Rompp, R. V. van Nieuwpoort, and H. E. Bal (2007). Parallel processing of large datasets from nanolc-fticr-ms measurements. *Journal of The American Society for Mass Spectrometry 18*, 152–161.

van der Laan, M., E. Polley, and A. Hubbard (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology 6*(25).

Verbeke, G. and G. Molenberghs (2000). *Linear Mixed Models for Longitudinal Data*. Springer.

Vickers, A. and E. Elkin (2006). Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making 26*, 565–574.

Wolpert, D. (1992). Stacked generalization. *Neural Networks 5*, 241–259.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society 67*, 301–320.