Daniel YEKUTIELI

Department of Statistics and OR
School of Mathematical Sciences
Tel Aviv University

Research Statement 2013 - 2026

1 Overview

In my MSc and PhD theses and the first part of my career I studied and developed frequentist FDR controlling methodology [1], [2], [3]. Working on False Coverage-statement Rate (FCR) controlling confidence intervals [4] and hierarchical FDR controlling methodology [5], I came to the realization that the frequentist approach may not offer viable solutions for more complicated multivariate selective inference problems. In [6] I studied the effect of selection on Bayesian inference and presented a Bayesian framework for selective inference. After writing this paper and working on empirical Bayes (eBayes) replicability analysis with Prof. Ruth Heller in [17], I decided to change the focus of my research to developing Bayesian methodology. Initially, empirical Bayes eBayes methodology for controlling the FDR. In the last few years I apply Bayesian methods for general statistical analyses like fitting GLM and nonparametric density estimation.

2 Collaborations 2014 - 2025

In this section I discuss my work on collaborative projects that did not include development of Bayesian methods. [7], [8], [9] and [10] are case studies and application papers I published with colleagues and students I supervised. [11] presents a statistical parameter estimation and testing framework for radar target detection. The proposed algorithm is shown to detect targets by transmission of an unmodulated single pulse, in cases where classical methods based on matched filter fail.

[12] and [13] resulted from joint research with Prof. Osant Ashur-Fabian and Dr. Eilon Krashin, from the Meir Medical center, whose goal was to develop diagnostic and prognostic scores for pancreatic cancer from clinical data mined from Clallit Electronic Health Records for all patients diagnosed with pancreatic cancer between 2000-2018 as well as age and gender matched controls. Our research received funding from the Tel Aviv University AI and Data Science Center that was used to support Dr. Zvi Segal's work on his MSc. thesis in Statistics.

[14] is joint work with Dr. Asaf Weinstein, from the Hebrew University, that presents FCR controlling confidence intervals for selected parameters that determine the parameter's sign (by consisting of either positive or negative parameter values) while balancing between the power to make more discoveries and the length of the confidence intervals.

For the last two years I am working with Dr. Baruch Ziv and Prof. Hadas Saaroni on producing an updated Typical Meteorological Year (TMY) for Israel. TMY is a collation of weather data for specific locations of representative hourly values of solar radiation and meteorological elements for a one-year period. The TMY is generated from 18 years of weather records of the Israeli Meteorological

Service and is used in simulations assessing expected heating and cooling costs for designing buildings and energy systems. Our project is funded by a grant from the Israeli Ministry of Energy and Infrastructure that is supporting Ms. Adi Giat on her MSc. thesis.

3 Implementation of Bayesian methodology 2015 - 2017

[15] is a methodological paper suggesting a Bayesian framework for constructing tests for complex hypotheses on cross-tabulated data for cases where a simple test statistic is not available. The tests are Bayesian extensions of the likelihood ratio test, they are also closely related to Bayes factors and Bayesian FDR controlling testing procedures. The motivating example is constructing a test for discovering Simpson's Paradox.

[16] presents a new algorithm for finding coherent and flexible modules in three-way data. The algorithm is a hierarchical Bayesian (hBayes) data model implemented with a Gibbs sampler. It outperformed extant methods on simulated and on real data. It was successfully used to dissect key components of septic shock response from time series measurements of gene expression and detect patient-specific module augmentations that were informative for disease outcome. it detected the pertinent brain region activity from functional magnetic resonance imaging time series of subjects at rest.

4 eBayes GWAS replicability 2014 - 2017

Genome-Wide Association Studies (GWAS) are performed to discover Single Nucleotide Polymorphisms (SNP) that are associated with an outcome of interest. The goal of replicability analysis is to discover SNP associations that are present in more than one study when combining information from several GWAS. [17] phrase replicability analysis as a Bayesian multiple group mixture model and present an eBayes method for controlling the FDR for replicability discoveries. The eBayes approach derives Bayes rules for classification that – per construction – produce FDR controlling tests with maximal power. [17] further illustrate that the BH procedure has very little power for discovering replicability because the no-replicability null is a composite null hypothesis and the corresponding p-value is a suboptimal summary for the multivariate statistics. [18] presents the R *repfdr* package that implements the eBayes approach for replicability analysis and meta-analysis. Our research on replicability was supported by a grant from the Israeli Science Foundation.

[19] presents a powerful big-data implementation of eBayes replicability analysis. The SCREEN (Scalable Cluster-based REplicability ENhancement) algorithm consists of (1) clustering an estimated correlation network of the studies, (2) learning replicability (e.g., of genes) within clusters, and (3) merging the results across the clusters using dynamic programming. Applied to a collection of 29 case-control large-scale gene expression cancer studies, SCREEN detected a large up-regulated module of genes related to proliferation and cell cycle regulation. On a pan-cancer study that examined the expression profiles of patients with or without mutations in the HLA complex, SCREEN detected an active module of up-regulated genes related to immune responses that includes most of the genes reported in the original study, and many new genes that were needed to establish the

connectivity of the module.

5 Deconvolution analysis 2018 - 2025

In this section I describe work on statistical analyses of mixture models whose goal is to provide inferences regrading the mixing distribution. For motivation, consider a meta-analysis with substantial unexplained between-study heterogeneity. As GRADE criteria highlight, unexplained heterogeneity reduces certainty in the evidence, resulting in limited confidence in effect estimates.

[20] presents a new clinically useful approach to assessing intervention effects in normal random-effects meta-analyses with strong unexplained heterogeneity. By constructing confidence intervals for the CDF of the between-study effect distribution we provide estimates and uncertainty assessments for the effect in a new group of patients. In one example, our method illustrated that evidence from a meta-analysis did not support authors' highly publicized conclusion that hypericum is as effective as other antidepressants. In the second example, our method provided insight into a subgroup analysis of the effect of ribavirin in hepatitis C, demonstrating clear important benefit in one subgroup but not in others. As part of his MSc. thesis, David Grabois is preparing a R package implementing the methods developed for [20] and useful model diagnostics, which we plan to publish shortly with additional illustrative case studies.

In [21] we present methodology for constructing pointwise confidence intervals for the CDF and quantiles of the mixing distributions from binomial mixture distribution samples. No assumptions are made on the shape of the mixing distribution. The confidence intervals are constructed by inverting exact tests of composite null hypotheses regarding the mixing distribution. The method may be applied to any deconvolution approach that produces test statistics whose distribution is stochastically monotone for stochastic increase of the mixing distribution. We propose a hBayes approach, which uses Finite Polya Trees (FPT) for modeling the mixing distribution, that provides stable and accurate deconvolution estimates without the need for additional tuning parameters. Our main technical result establishes the stochastic monotonicity property of the test statistics produced by the hBayes approach. Leveraging the need for the stochastic monotonicity property, we explicitly derive the smallest asymptotic confidence intervals that may be constructed with our method. Raising the question whether it is possible to construct smaller confidence intervals for the mixing distribution without making parametric assumptions on its shape. In the R *mcleod* package Dr. Barak Brill implemented the hBayes modeling approach for Binomial and Poisson mixture distribution samples and performs the Binomial and Poisson regression modeling with nonparametric random intercept terms that Dr. Brill developed in PhD thesis [22].

6 hBayes for risk minimization 2019 -

In this line of work we use hBayes modeling to approximate Bayes rules in invariant decision problems. Invariant decision problems are statistical problems in which the problem's structure, including the underlying distributions and loss function, remains unchanged under certain groups of transformations. In invariant decision problems the risk of invariant actions is the same for all parameter

values in orbits of the group of transformation. Thus Bayes rules that use the Haar measure of the group of transformation as the prior distribution are also invariant actions and they minimize the risk for any fixed unknown parameter value. The idea of using Bayes rules for minimizing frequentist risk was suggested by Herbert Robbins in the early 1950's and invariant decision problems were widely studied in the 1960's.

We use invariance principles in high-dimensional problems in which the risk for the Bayes rule on the orbit is considerably smaller than the risk of existing methods (which are typically also invariant under the group of transformations). The main inferential problem is finding the orbit of the unknown parameter value. The Bayes rule is then derived by sampling the posterior distribution with respect to the Haar measure prior on the orbit.

In [23] we apply invariance principles for constructing optimal regularized estimators in a given GLM with fixed effects. We present an algorithm for sampling posterior distributions for a hBayes model that uses FPT to generate the prior distribution of the model coefficients and the uniform prior on model coefficient permutations. We demonstrate in examples that the posterior mean under the postulated model adapts non-parametrically to the empirical CDF of the true coefficients. Correspondingly, the Bayes rules for the hBayes approach are very similar to Bayes rules for the prior distribution that randomly permutes the true model coefficients. Numerical experiments show that our method has better estimation and prediction accuracy compared to various parametric and nonparametric alternatives, from relatively standard L_p-regularized estimators to modern penalized-likelihood and Bayesian estimators for high dimensional regression.

I am currently working with Prof. Malgorzata Bogdan and Jonas Wallin and Dr. Daniel Xiang on implementing this inferential approach to estimate the covariance matrix from multivariate normal samples, for which we consider invariance under the group of orthogonal matrices and orbits consist of covariance matrices with the same ordered eigenvalue vectors.

7 hBayes framework for Binary Expansion Statistics 2025 -

In this section I briefly describe planned research with Prof. Kai Zhang and Jan Hannig from the University of North Carolina, which will be supported in the next 3 years by a grant from the Binational Science Foundation. We plan to use the hBayes approach for nonparametric density estimation presented in [24] for performing the binary expansion testing framework and binary expansion linear effect modeling developed by Prof. Zhang. An important part of our project will be to use the expertise of Prof. Hannig to formulate precise probabilistic statements, based on the generalized fiducial distribution, for the inferences produced by the hBayes models.

References

[1] **D. Yekutieli**, Y. Benjamini "Resampling based False Discovery Rate controlling multiple testing procedure for correlated test statistics". J Statist. Plann. Inference 82 (1999), 171-196.

[2] **D. Yekutieli**, Y. Benjamini "The control of the False Discovery Rate in multiple testing under dependency". Annals of Statistics 29(4) (2001), 1165-1188.

[3] Benjamini, Y., Krieger, A.M., **Yekutieli, D.** "Adaptive Linear Step-up False Discovery Rate controlling procedures" Biometrika (3): 491-507 Sep 2006.

[4] Y. Benjamini, **D. Yekutieli** "False Discovery Rate controlling confidence intervals for selected parameters" Journal of the American Statistics Association 100, 469 (2005), 71-80.

[5] **Yekutieli, D.,** "Hierarchical False Discovery Rate controlling methodology" Journal of the American Statistical Association, 2008, 103 (481) 309-316

[6] **Yekutieli D.** "Adjusted Bayesian inference for selected parameters" *Journal of the Royal Statistical Society: Series B*, 2012, **74** (3) 515 - 541.

[7] Gottesman, T., Yossepowitch, O., Lerner E., , Schwartz-Harari, O., Soroksky, A., **Yekutieli, D.**, Dan, M., (2014) "The accuracy of Gram stain of respiratory specimens in excluding Staphylococcus aureus in ventilator-associated pneumonia." *Journal of critical care*, 2014, 29, no. 5 : 739-742.

[8] Adetayo K., Shkedy Z., Lin D., Van Sanden S., Cortinas Abrahantes J., Goehlmann H., Bijnens L., **Yekutieli D.**, Aerssens J., Camilleri M., Talloen W., "Translation of disease associated gene signatures across tissues." *International journal of data mining and bioinformatics* , 2015, 11.3, 301-313.

[9] Angelini C., Heller R., Volkinshtein R., **Yekutieli D.**, "Is this the right normalization? A diagnostic tool for ChIP-seq normalization." *BMC bioinformatics*, 2015, 16(1), 150.

[10] Daniels D., Guez D., Last D., Hoffmann C., Nass D., Talianski A., Tsarfaty G., Salomon S., Kanner A.A., Blumenthal D.T., Bokstein F., Harnof D., **Yekutieli D.**, Zamir S., Cohen Z.R., Zach L., Mardor Y., "Early Biomarkers from Conventional and Delayed-Contrast MRI to Predict the Response to Bevacizumab in Recurrent High-Grade Gliomas" *American Journal of Neuroradiology*, 2016, 37.11: 2003-2009.

[11] Ferdman Y., **Yekutieli D.**, Sochen N., "A method for radar detection and range-Doppler estimation," (2017) IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS), Tel-Aviv, 2017, pp. 1-6.

[12] Krashin E., Silverman B., Steinberg D.M., **Yekutieli D.**, Giveon S., Fabian O., Hercbergs A., Davis P.J., Ellis M., Ashur-Fabian O. (2021) "Pre-diagnosis thyroid hormone dysfunction is associated with cancer mortality." *Endocr Relat Cancer*, **28**:705-713

[13] Krashin E.; Silverman B., Steinberg D.M., **Yekutieli D.**, Giveon S., Fabian, O., Hercbergs A, Davis P.J.; Ellis M., Ashur-Fabian O., (2021) "Opposing effects of thyroid hormones on cancer risk: A population-based study." *Eur. J. Endocrinol.* **2021**, 184, 477-486.

[14] Weinstein, A., Yekutieli, (2020) "Selective sign-determining multiple confidence intervals with FCR control.", *Statistica Sinica*, **30** 531-555

[15] **Yekutieli D.** "Bayesian tests for composite alternative hypotheses in cross-tabulated data" *TEST*, 2015, 24, no. 2: 287-301.

[16] Amar D., **Yekutieli D.**, Maron-Katz A., Hendler T., Shamir R., "A hierarchical Bayesian model for flexible module discovery in three-way time series data" *Bioinformatics*, 2015, 31.12: i17-i26.

[17] Heller R., **Yekutieli D.**, "Bayesian FDR procedure for discovering replicability in Genome-Wide Association Scans," *The Annals of Applied Statistics* **8**, 2014, (1) 481-498.

[18] Heller R., Yaacoby S., **Yekutieli D.** "repfdr: a tool for replicability analysis for genome-wide association studies," *Bioinformatics*, 2014 : btu434.

[19] Amar, D., Shamir, R., **Yekutieli, D.**, "Extracting replicable associations across multiple studies: algorithms for controlling the false discovery rate." (2017). PLOS computational biology. *https://doi.org/10.1371/journal.pcbi.1005700*

[20] Saad A., **Yekutieli D.**, Lev-Ran S., Gross R., Guyatt G.H., "Getting More out of Meta-Analyses: A new approach to meta-analysis in light of unexplained heterogeneity." (2019). *Journal of Clinical Epidemiology,* Volume 107, Pages 101-106, *https://doi.org/10.1016/j.jclinepi.2018.11.023*

[21] Basu, P., Brill, B. and **Yekutieli, D.** (2025) "Exact Confidence Intervals for the Mixing Distribution from Binomial Mixture Distribution Samples" *Journal of Computational and Graphical Statistics* , October, 1–18. doi:10.1080/10618600.2025.2573147.

[22] Brill, B. (2022), "Statistical challenges in microbiome research and analysis of discrete compositional data", Ph.D. Dissertation, Tel Aviv University, https://tau.primo.exlibrisgroup.com/discovery/fulldisplay/alma9933425419004146/972TAU_INST:TAU

[23] Weinstein, A., Wallin, J., **Yekutieli**, D., Bogdan, M., (2023) Weinstein, A., Wallin, J., **Yekutieli, D.**, Bogdan, M. (2025) "Nonparametric shrinkage estimation in generalized linear models via Polya trees," *Statistica Sinica* **38** (4), 1-11.

[24] **Yekutieli, D.**, (2024) "Distribution-Free Bayesian multivariate predictive inference" *arXiv:2110.07361*