Volume 37
Number 3
September - December 2025

Statistica Applicata

ITALIAN JOURNAL OF APPLIED STATISTICS

ISSN: 1125-1964

FIRENZE UNIVERSITY PRESS

Statistica Applicata – ITALIAN JOURNAL OF APPLIED STATISTICS

EDITORIAL TEAM

EDITOR IN CHIEF
- Francesco Palumbo, Università di Napoli Federico II, Naples, Italy

CO-EDITORS ON A SPECIFIC SUBJECT
- Alessandro Celegato, AICQ Centronord - Quality and techonology in production
- Adriano Decarli, Università di Milano, IRCCS /INT Foundation, Milan, Italy - Social and health studies
- Luigi Fabbris, Università di Padova, Padua, Italy - Surveys and experiments
- Vittorio Frosini, Università Cattolica del Sacro Cuore, Milan, Italy - Book review
- Antonio Giusti, Università di Firenze, Florence, Italy - Data Science
- Paolo Mariani, Università di Milano Bicocca, Milan, Italy - Social and economic analysis and forecasting

SCIENTIFIC COMMITTEE
- Thomas Aluja, UPC, Barcelona, Spain
- Paul P. Biemer, RTI and IRSS, Chicago, USA
- Jörg Blasius, Universität Bonn, Bonn, Germany
- Irene D'Epifanio, Universitat Jaume I, Castelló de la Plana, Spain
- Vincenzo Esposito Vinzi, ESSEC Paris, France
- Gabriella Grassia, Università di Napoli Federico II, Naples, Italy
- Michael J. Greenacre, UPF, Barcelona, Spain
- Salvatore Ingrassia, Università di Catania, Catania, Italy
- Ron S. Kenett, KPA Ltd. and Samuel Neaman Institute, Technion, Haifa, Israel
- Stefania Mignani, Università di Bologna Alma Mater, Bologna, Italy
- Tormod Naes, NOFIMA, Oslo, Norway
- Alessandra Petrucci, Università di Firenze, Florence, Italy
- Monica Pratesi, Università di Pisa, Pisa, Italy
- Maurizio Vichi, Sapienza Università di Roma, Rome, Italy
- Giorgio Vittadini, Università di Milano Bicocca, Milan, Italy
- Adalbert Wilhelm, Jacob University, Breimen, Germany

ASSOCIATE EDITORS
- Francesca Bassi, Università di Padova, Padua, Italy
- Bruno Bertaccini, Università di Firenze, Florence, Italy
- Matilde Bini, Università Europea, Rome, Italy
- Giovanna Boccuzzo, Università di Padova, Padua, Italy
- Maurizio Carpita, Università di Brescia, Brescia, Italy
- Giuliana Coccia, ISTAT, Rome, Italy
- Fabio Crescenzi, ISTAT, Rome, Italy
- Franca Crippa, Università di Milano Bicocca, Milan, Italy
- Corrado Crocetta, Università di Foggia, Foggia, Italy
- Cristina Davino, Università di Napoli Federico II, Naples, Italy
- Loretta Degan, Gruppo Galgano, Milan, Italy
- Tonio Di Battista, Università di Chieti-Pescara "Gabriele D'Annunzio", Pescara, Italy
- Tommaso Di Fonzo, Università di Padova, Padua, Italy
- Francesca Di Iorio, Università di Napoli Federico II, Naples, Italy
- Simone Di Zio, Università di Chieti-Pescara "Gabriele D'Annunzio", Pescara, Italy
- Filippo Domma, Università della Calabria, Rende, Italy
- Alessandra Durio, Università di Torino, Turin, Italy
- Monica Ferraroni, Università di Milano, Milan, Italy
- Giuseppe Giordano, Università di Salerno, Salerno, Italy
- Michela Gnaldi, Università di Perugia, Perugia, Italy
- Domenica Fioredistella Iezzi, Università di Roma Tor Vergata, Rome, Italy
- Michele Lalla, Università di Modena e Reggio Emilia, Modena, Italy
- Maria Cristina Martini, Università di Modena e Reggio Emilia, Modena, Italy
- Fulvia Mecatti, Università di Milano Bicocca, Milan, Italy
- Sonia Migliorati, Università di Milano Bicocca, Milan, Italy
- Michelangelo Misuraca, Università della Calabria, Rende, Italy
- Francesco Mola, Università di Cagliari, Cagliari, Italy
- Roberto Monducci, ISTAT, Rome, Italy
- Isabella Morlini, Università di Modena e Reggio Emilia, Modena, Italy
- Biagio Palumbo, Università di Napoli Federico II, Naples, Italy
- Alfonso Piscitelli, Università di Napoli Federico II, Naples, Italy
- Antonio Punzo, Università di Catania, Catania, Italy
- Silvia Salini, Università di Milano, Milan, Italy
- Luigi Salmaso, Università di Padova, Padua, Italy
- Germana Scepi, Università di Napoli Federico II, Naples, Italy
- Giorgio Tassinari, Università di Bologna Alma Mater, Bologna, Italy
- Ernesto Toma, Università di Bari, Bari, Italy

- Rosanna Verde, Università della Campania "Luigi Vanvitelli", Caserta, Italy
- Grazia Vicario, Politecnico di Torino, Turin, Italy
- Maria Prosperina Vitale, Università di Salerno, Salerno, Italy
- Susanna Zaccarin, Università di Trieste, Trieste, Italy
- Emma Zavarrone, IULM Milano, Milan, Italy

EDITORIAL MANAGER
- Domenico Vistocco, Università di Napoli Federico II, Naples, Italy

EDITORIAL STAFF
- Antonio Balzanella, Università della Campania "Luigi Vanvitelli", Caserta, Italy
- Luca Bagnato, Università Cattolica del Sacro Cuore, Milan, Italy
- Paolo Berta, Università di Milano Bicocca, Milan, Italy
- Francesca Giambona, Università di Firenze, Florence, Italy
- Rosaria Romano, Università di Napoli Federico II, Naples, Italy
- Rosaria Simone, Università di Napoli Federico II, Naples, Italy
- Maria Spano, Università di Napoli Federico II, Naples, Italy

A.S.A CONTACTS
Principal Contact
Francesco Palumbo (Editor in Chief)
editor@sa-ijas.org

Support Contact
Domenico Vistocco (Editorial Manager)
ijas@sa-ijas.org

JOURNAL WEBPAGE
https://www.sa-ijas.org/ojs/index.php/sa-ijas

Statistica Applicata – Italian Journal of Applied Statistics is a four-monthly journal published by the Associazione per la Statistica Applicata (A.S.A.), Largo Gemelli 1 – 20123 Milano, Italy (phone + 39 02 72342904). Advertising:Cleup sc, via G. Belzoni, 118/3 – 35128 Padova, Italy (phone +39 049 8753496 – Fax +39 049 9865390), email: info@cleup.it.
Rules for manuscript submission: https://www.sa-ijas.org/ojs/index.php/sa-ijas/about/submissions
Subscription: yearly €103.30; single copy €40.00; A.S.A. associates €60.00; supporting institutions: €350.00. Advertisement lower than 70%. Postal subscription Group IV, Milan. Forum licence n. 782/89. CLEUP SC on behalf of ASA, 7 March 2023.

Statistica Applicata – Italian Journal of Applied Statistics is associated to the following Italian and international journals:

QTQM – Quality Technology & Quantitative Management (http/web.it.nctu.ew/~qtqm/)

SINERGIE – Italian Journal of Management

Statistica Applicata – Italian Journal of Applied Statistics (ISSN:1125-1964, E-ISSN:2038-5587) applies the Creative Commons Attribution (CC BY) license to everything we publish.



Pubblished: September 2025

© 2025 Author(s)
Content license: except where otherwise noted, the present work is released under Creative Commons Attribution 4.0 International license (CC BY 4.0: https://creativecommons.org/licenses/by/4.0/legalcode). This license allows you to share any part of the work by any means and format, modify it for any purpose, including commercial, as long as appropriate credit is given to the author, any changes made to the work are indicated and a URL link is provided to the license.
Metadata license: all the metadata are released under the Public Domain Dedication license (CC0 1.0 Universal: https://creativecommons.org/publicdomain/zero/1.0/legalcode).

Published by Firenze University Press
Powered by Firenze University Press

Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

Summary

Statistica Applicata
ITALIAN JOURNAL OF APPLIED STATISTICS
Vol. 37, Number 3

249 Bazzoli, M., Marzadro, S. — *Matching and integration of registry and survey data on the dynamics of work histories: A pilot study*

273 Adeniyi, I.A., Shobanke, D.A. — *A modified correlation based regularization technique for regression estimation and feature selection*

297 Satheesh Kumar, C., Harisankar, S. — *On a wide class of generalized geometric distribution for over and under dispersed data sets*

317 Domagni, F., Neubauer, M. — *Non-Uniqueness of E(s 2)-Optimal Supersaturated Designs for N Application to the Case N = 10 Runs ≡ 2 (mod 4) Runs with*

doi.org/10.362653/IJAS.16751

MATCHING AND INTEGRATION OF REGISTRY AND SURVEY DATA ON THE DYNAMICS OF WORK HISTORIES: A PILOT STUDY

Martina Bazzoli[1], Sonia Marzadro
FBK-IRVAPP, Trento, Italy

Abstract. *In the last decade the practice of combining longitudinal surveys with administrative data has been affirmed, enabling the integration of the complementary advantages offered by these two data sources. In this perspective, the paper describes a pilot study conducted in 2015 involving deterministic matching and integration between a retrospective panel survey carried out on a representative sample of households living in the province of Trento (Italy) and the register data from the provincial section of the administrative archive of INPS (the Italian social security agency). The aim was to create a comprehensive database for the study of work histories. Through the survey we address the main limitation of INPS data, which do not cover the universe of workers and jobseekers. Conversely, the administrative data, providing richer and more reliable information on most work episodes and instances of subsidized unemployment, help mitigate one of the most significant sources of errors in the panel surveys, i.e. the distortions caused by memory bias.*

Keywords: *data integration, administrative data, survey data, labour market, work histories*

[1] Corresponding author: Martina Bazzoli, email: bazzoli@irvapp.it
The research has been carried out in the framework of the Agreement Protocol between FBK-Irvapp, the Trentino-Alto Adige Regional Directorate of Inps and Ispat (n. 157/176704109). We are grateful to Antonio Schizzerotto and Ugo Trivellato who coordinated the research project. We also thank Marco Zanotelli and Giovanna Fambri for the support provided by Inps and Ispat respectively in data acquisition and integration, and to Bruno Anastasia, Cinzia Giannelli, Gianluca Mazzarella and Michele Mosca for their comments and suggestions.

1. INTRODUCTION

Are today's youth worse off than yesterday's? The question often arises in the press, in public debate and in economic and social research. However, answering this question is difficult. Indeed, being worse – or better or in the same way – involves comparisons between generations. Such comparisons require dealing with complex methodological issues, but even more, require longitudinal datasets following different cohorts of individuals – today's young people and those of yesterday – for their entire life course or at least for a long part of it.

Usually, many countries collect this kind of data through large multi-purpose household panel surveys using appropriate research designs. In this regard, Understanding Society is the exemplary longitudinal study of the United Kingdom launched in 2009 with annual waves on individuals of about 40,000 families.[2]

In countries - typically Scandinavian - where the production of statistical information is largely based on registers of administrative origin containing comprehensive information on the population or a key sub-population, the alternative to survey data is made up of longitudinal databases for thematic areas derived directly from registers.[3]

Rather than choosing one data source over another, the practice of combining longitudinal surveys with administrative data in such a way as to maximize the strengths and minimize the weaknesses of each source – a process called *data integration* – has been established in the last decade especially in English-speaking countries (Calderwood and Lessof, 2009). The example is still provided by Understanding Society, in which, with the consent of the adult respondents, the pertinent administrative data on health and education are linked to the survey data.

[2] The sample also includes 8,000 households from the previous British Household Panel Survey, which started in 1991. Among other longitudinal household surveys, is worth mentioning the German Socio-Economic Panel (SOEP), conducted since 1994, and the US Panel Study of Income Dynamics (PSID), started in 1968.

[3] Here the Danish experience in the labour market field is paradigmatic: Statistics Denmark has developed and updated annually, IDA (Integrated Database for Labour Market Research), a base of longitudinal data on the population of workers and enterprises since 1980.

Unfortunately, in Italy, longitudinal databases that follow an appropriate group of individuals over a long-time, whether achieved through household panel surveys[4] or derived from administrative sources[5], are not currently available at the national level despite the considerable progress made by the National Statistical Institute (Istat) in terms of integration of administrative and survey data[6]. Moreover, it must be noted that Italy suffers from restrictive legislation on privacy and protection of personal data which limits the process of integrating data for scientific purposes (Trivellato, 2019). Among the few notable Italian experiences, the AD-SILC[7] database stands out as a significant contribution. It was constructed by longitudinally linking administrative data from various National Social Security Institute archives, covering diverse social groups such as public and private sector workers, self-employed professionals, pensioners, and recipients of other insurance-based welfare benefits. These data were further integrated with microdata from IT-SILC, the Italian version of the EU-SILC survey.

[4] The large national cross-sectional surveys with a longitudinal component, IT-SILC (the Italian section of EU-Survey on Income and Living Conditions), the Labour Force Survey and the Survey on Household Income and Wealth of the Bank of Italy cover individuals for a short period of time. Then there are two other multi-purpose household panel surveys but they were conducted between the end of the twentieth and the beginning of the twenty-first century: the European Community Household Panel in which also Italy were included (8 annual waves between 1994 and 2001, and the Italian Longitudinal Household panel (5 biennial waves, from 1997 to 2005).

[5] Administrative data are restricted to segments of population. The most interesting experience in the labour market field concerns longitudinal databases that have been built on management archives of the National Social Security Institute. Among them the WHIP (Work Histories Italian Panel) contains a linked employer-employee longitudinal database. WHIP has been widely used for studies on the mobility of the Italian labour market however his update has stopped in 2004.

[6] Such dataset is, indeed, mainly focused on firms and income variables (Sestito et Trivellato, 2011; Istat, 2009

7 In the first version of AD-SILC, INPS data were merged with the 2015 IT-SILC microdata (Fondazione G. Brodolini, MEF, 2016). In the latest version (Conti et al., 2023), AD-SILC integrates longitudinal data from IT-SILC (covering the period from 2004 to 2017) and INPS data. AD-SILC has been used primarily, but not exclusively (Naticchioni et al., 2016, Fabrizi and Rocca, 2024), as the basis for the dynamic microsimulation model T-DYMM developed by the Treasury Department of the Italian Ministry of Economy and Finance (Conti et al., 2023).

Given this context, in 2015 FBK-IRVAPP, in collaboration with the Regional Directorate of the National Social Security Institute (INPS) and the Institute of Statistics of the Province of Trento (Ispat), conducted a research project aiming at providing a picture, as accurately as possible, of the work career of a sample of individuals living in Trentino. This was achieved through the deterministic matching and integration of a panel survey on households' life conditions (*Indagine sulle condizioni di vita delle famiglie trentine*, hereafter ICFT) carried out on a representative sample of about 2,500 households and register data on all employees managed by the local agency of the National Social Security Institute (hereafter INPS).

The study was conducted on a small-scale context - the autonomous province of Trento - and faced with empirical research issues not based on a well-grounded literature. However, in our opinion, this study holds significant value as one of the rare Italian examples of data integration. It is, as far as we know, the only one that not only merges but fully integrates the employment episodes recorded in administrative data (at that time) with the entire work histories collected through a retrospective survey thereby leveraging the strengths of both sources to provide a more comprehensive and accurate representation of individuals' employment trajectories.

The aim of this paper is to describe the archives involved, the integration process, the various challenges faced, and the characteristics of the resulting dataset. This dataset has been used to explore the main features of the evolution of the work histories of individuals over a time span that extends, overall, from 1974 to 2014. Although this part is not discussed here, a brief summary of the empirical evidence is provided in Section 5 (see Bazzoli et al., 2018 for more detailed findings).

The paper is structured as follows. In Section 2 we provide an overview of important considerations for linking administrative and survey data for research. The two sources of data used are introduced in Section 3, while Section 4 describes the integration procedures. In Section 5, we recall the key findings on the employment dynamics of young people over the last 40 years as studied using the aforementioned database. Conclusions are offered in Section 6.

2. INTEGRATING ADMINISTRATIVE AND SURVEY DATA TO BUILD EVIDENCE

Data sources are an essential part of research and whether administrative and survey data differ significantly is becoming a matter of discussion among researchers. In addition to the tightening of survey budgets, it is well-known that falling response rates and memory recall errors are the two main challenges that threaten the quality of collected survey data thus reducing the confidence in the conclusion drawn from such data. Moreover, there is an increasing need to satisfy a growing demand from users for good quality statistics and above all enabling faster measurement of new phenomena.

Administrative data can be a powerful resource in this respect particularly because of the insights these data might offer into social inequality, human behaviour and the effectiveness of social policies (Einav and Levin, 2014; Connelly 2016; Card et al., 2010). Moreover, they are an authoritative source of data, cover a large number of observations, are regularly updated and thus are available for long periods of time (Künn, 2015). Therefore, they can be used not only to enrich survey data, but also to lower survey costs by requiring fewer questions during interviews, reduce the burden on respondents and lessen survey dropout and item nonresponse rates. As a matter of fact, linking administrative data to survey data has become a promising and innovative strategy which affects the quality and quantity of research and increases the potential of data (Crato and Paruolo, 2019; Künn, 2015; Al Baghal et al., 2014; Calderwood and Lessof, 2009).

However, despite the clear advantages, there are a number of challenges in using and integrating that kind of data. First of all, there are ethical and legal considerations that need to be resolved. Indeed, legal constraints on the use of data and the need for anonymization restrict their access and content (Livraga, 2019; Trivellato, 2019; Sakshaug et al., 2012). Moreover, requesting consent to use the linked data may introduce consent bias (that might occur when consenters differ from non-consenters) or reduce response rates, introducing selection bias (Sakshaug et al., 2012).

Another issue deals with the time and conceptual alignment. A reference period in survey data is commonly the state at the time of the interview (e.g. current activity status) or is collected via retrospective questions (e.g. career history) while the data in administrative archives usually are registered in a fixed

point in time. Finally, variables collected for administrative purposes sometimes may diverge from fundamental economic concepts (e.g. unemployment status) or are missing because they are not of interest to the administrators who collect data (such as attitudes and family relationships).

The linkage of register variables to surveys is quite straightforward if the survey units also have valid identifiers as in the case of survey population drawn from an administrative archive. In this case the data can be directly linked via exact/deterministic matching because both data sources contain a unique personal identifier. In the absence of unique identifiers, statistical matching or modelling approaches have to be used instead (Harron et al., 2017).

Although the primary motivation for linking to administrative data is usually to enhance survey data, this can be done in different ways because it may involve pooling or combining information. On one side, administrative data are used to supplement survey data. In this case they are used as a direct substitute for survey questions (low integration). On the other hand, administrative data are used to complement/validate survey data. This occurs, for example, when administrative data is available only for a subgroup of survey respondents (high integration), like in the present case.

In the next sections we clarify how we have used the two sources of data mentioned before – a multi-purpose panel survey with retrospective part on working life of a representative sample of individuals and register data that provide much more detailed and reliable information on work histories than survey data but on a subsample of workers – in order to create a database that enables to the study of individuals' work histories over a period of nearly four decades.

3. DATA DESCRIPTION

3.1 THE ICFT PANEL SURVEY

The survey used is a multi-purpose longitudinal survey with a retrospective part on working life conducted on a representative sample of about 2,500 households from Trentino, an autonomous province of Northern Italy. It was started in 2005/06 with biennial waves[8] conducted using face-to-face computer assisted personal interviewing (CAPI).

[8] For details see Fambri and Schizzerotto, 2008.

The first wave gathered retrospective information on all significant events occurring to the members of the sample in the period between their births and the date of the interview. The purpose of each of the subsequent surveys was to update this information, recording all significant events of life histories that occurred to the members of the sample in the period between the previous interview and the date of the current one.

The most complex part of the interview was the reconstruction of the individual working career, including any periods of unemployment or inactivity. The survey gathered the dates of each episode experienced by the interviewees, as well as the main features (such as labour contract, activity sector, type of work). Each 'work-history' studied is represented by a continuous sequence - of varying duration – of distinct episodes, each at least one-month long. The starting date of each episode (except the first) corresponds to the ending date of the one that precedes it. The career of a respondent is thus represented by a sequence of non-overlapping episodes - of work, unemployment or inactivity - lasting at least one month between the date of the first entry into the labour market and the date of the interview.

The sample for the experience of integration presented in this paper is composed of 5,756 adults interviewed during the 2012 wave.

3.2 THE ADMINISTRATIVE INPS DATA

For each individual belonging to the reference sample, individual information was then acquired in 2015 from the National Social Security Institution (INPS), from "*Cassetto previdenziale del cittadino*", which contains the list of contributions recorded in favour of the worker from the opening of an insurance position.

Contributions belong to different funds: 1) private sector employees; 2) clergy; 3) showbusiness and entertainment; 4) separate management which includes all atypical workers. Unfortunately, public employees[9], professionals registered with their own pension fund and workers employed in the informal economy are not included in the archive. For each spell, INPS collects starting and ending dates and, for private sector employees only, some main features such

[9] The INPDAP, the National Institute for Social Security and Assistance for Public Administration Employees, was abolished, and its functions were transferred to the INPS as of January 2012. At the time of extracting the INPS data (2015), information on public sector employees was not yet available.

as the type of work, salary, weeks of contributions paid. It should be noted that for all contributions belonging to the separate management, starting and ending dates are conventionally fixed at the 1st January and the 31st December respectively. Since 1974, for most of the employees, there is also a unique number identifying the company (*matricola*) that made it possible to link, for each contribution spell, the fiscal code of the company, its economic activity code, address, date of establishment and (potential) closure.[10]

In addition to the contributions paid by employers in the case of employees or by the self-employed themselves, the archive managed by INPS also includes non-working spells called imputed contributions (*contribuzioni figurative*) that include solidarity contracts, unemployment and mobility benefits.

4. THE INTEGRATION PROCESS

4.1 RECORD LINKAGE

The aim of record linkage was to identify pairs of records in the two sources using the individual fiscal code as the matching key.

We successfully matched all the individual records of the wave 2012 of the ICFT survey (N=5,733) with INPS data (Table 1). The result was a combined database of 5,489 adults with at least one working episode, in the survey and/or in the register. The vast majority of the sample (4,551 subjects) was in both databases. Then, 507 individuals in the survey were not found in INPS database since their jobs were not recorded, by definition, in the administrative data and 431 subjects didn't declare any job experience in the survey even though they had one. Finally, 244 subjects were excluded because they didn't have any work episodes.

[10] These data come from the E-Mens flows.

Table 1: Sampling of individuals after the matching

	N
Individuals over 18s interviewed in ICFT survey (wave 2012) [a]	5,733
• With at least one working episode in ICFT survey	
o and at least one episode in *INPS*	4,551
o but not in *INPS*	507
• Without any working episode in ICFT survey	
o But at least one episode in *INPS*	431
o And any working episode in *INPS*	244

[a] Net of 23 individuals with unknown fiscal code.

4.2 DATA PROCESSING

The next step was to integrate the information contained in the two databases. However, the study confronted us with the dual need to address different issues of definition and to develop multiple procedures for handling the data.

In particular, while the ICFT was explicitly designed to study work stories, processing register data was quite demanding as the archive meets the Institute's management purposes and is therefore not immediately usable by researchers. The register included about 158,000 records, 124,000 of which were work records and the rest were non-work records. In order to reconstruct the working life of each individual in the sample, the following steps have been followed.

First of all, the type of contribution was used to identify work episodes and, in particular, to classify all records belonging to a fund under the following employment conditions: "Employees in the private non-agricultural sector", "Artisans, retailers and entrepreneurs", "Freelance professionals", "Temporary workers and other self-employed individuals", "Agricultural workers (employees and self- employers)". Within the periods of imputed contributions, the time segments of unemployment benefits in the strict sense (ordinary unemployment benefits, mobility benefits, ASpI and MiniASpI) were identified[11].

[11] Register data does not provide complete information on non-employment spells but just on periods of subsidized unemployment which may include subsidised unemployment in the strict sense (ordinary unemployment benefit, mobility allowance, ASpI and MiniASpI) and other subsidised non-employment (reduced unemployment,

In order to have adequate statistical data to represent the work histories, it was necessary to carry out a meticulous activity of refinement of the administrative information since, according to the law, the same episode (both of employment and subsidized unemployment) could appear in the archive in more than one record. In particular, we addressed three main aspects:

i) duplicate episodes or events with no contributing weeks;
ii) segmented episodes;
iii) episodes of spurious mobility.

With the first cleaning procedure we ignored all those episodes, both of work and non-work, which, for administrative reasons, were *duplicates* or had *zero contribution weeks* (overall about 4,000 records). In particular, the former operation concerned episodes of non-agricultural employment bearing the identification of the company[12], self-employment (traders, craftsmen, entrepreneurs), agriculture (self-employed and employees) and episodes of subsidised unemployment stricto sensu. In the case of coincidence between start and end dates and, for work episodes, type of contribution and company, only one episode was maintained.

The second intervention concerned the *segmented work episodes*, i.e. the (quite frequent) cases in which a working episode within the same company was recorded as if it were a plurality of different episodes with different but consecutive starting and ending dates.[13] In this case, the records were combined into a single work episode. The number of records was reduced by approximately 80,000 cases.

Finally, we detected the firms' legal transformations to handle spurious movements of workers. As is well known, firms can undergo changes over time: in their identifying characteristics (name or address), structural characteristics (size or prevailing activity) or in cases of mergers, demergers or deaths. In many cases, these changes do not result from real changes but from the adaptation of legal entities to laws and administrative rules. However, all these events, even

agricultural and construction). The latter have not been considered here because, conventionally, annual periods starting in January and ending in December are reported, thus overestimating the actual duration; moreover, the monetary transfer is also compatible with the inactive condition.

[12] No intervention has been carried out on non-agricultural employees with missing information on the company.

[13] Not necessarily from January to December.

those which, in fact, did not generate real discontinuities in the life of companies, were recorded in the administrative archives generating the so-called spurious flows.

For the study of work histories, it was, therefore, crucial to discriminate between movements originating from choices of workers and spurious movements that were induced by changes that had affected companies. The possibility provided by the administrative archive to observe both the individual employees and the companies for which they work, enabled us to deduce the existence of other underlying legal transformations from the simultaneous flow of workers between two companies, in order to discriminate between real movements (i.e. resulting from workers' decisions to change jobs) and spurious movements.[14] If we had ignored the specific nature of these flows, we would have observed a fictitious variation in the number of jobs.

In order to detect the presence of firms' legal transformations we proceeded as follows. Firstly, we considered all the episodes of private sector employees belonging to the regular fund (about 25,000). Then we identified the direct movements from one firm to another (the so-called job-to-job) and then we selected those that occurred over a period of 30 days (about 5,000).

An ad hoc employee/employers' database was then created containing, for each individual with at least one job-to-job within 30 days, the characteristics of the starting company and those of the ending one. Two criteria were then used to detect spurious moves:

i) *Similarity of attributes*: a link was established between two companies with different identification code (matricola) according to the similarity in the values of certain characteristics and, in particular, their *name* and *address*;

ii) Identification of firms' *legal transformations* through workers' movements: the existence of transformations was deduced from the flow of employees.

As regards the first criterion, movements between companies with the same *name* were initially candidates to be considered as spurious. This was the case

[14] The identification of spurious mobility relationships could only be carried out for employees with company identification. Therefore, in addition to working episodes prior to 1975, all episodes that fell under the separate management, in showbusiness and entertainment and clergy (as well as those of self-employment) were excluded. Given their peculiarity, episodes of temporary agency work of usual seasonal work were also excluded. Altogether we ignored from the analysis of spurious mobility about 45,000 episodes.

for companies with several operating units (e.g. Unicredit). Overall, 206 job-to-jobs have been initially identified using this criterion.

In addition to the name, a second characteristic that was used to establish a link between two companies was the *address*. However, considering the fact that the same address can be written in different ways (Via S. Croce 77, Trento vs. Via Santa Croce 77, Trento), we preferred to use geocoding via Google Maps. Based on the consistency between the geographical coordinates, the economic activity code and the difference between the date of cessation of the company of origin and the date of establishment of the company of arrival (delta), the following levels of quality of the combination have been defined:

- High level (5): companies with the same address, economic activity code equal (to the 3rd digit) and delta less than 30 days.
- Medium-high level (4): level 5 without an economic activity constraint.
- Medium level (3): level 5 but with a maximum discrepancy in the geographical coordinates of the two companies of 0.0001 points (corresponding to about half a block: an example is a building with several house numbers).
- Medium-low level (2): level 3 without constraint of economic activity
- Low level (1): maximum discrepancy in the geographical coordinates of the two companies of 0.001 points (about one block), economic activity code at the 2nd digit, time difference < 30 days.

The same ranking was also assigned when the date of cessation of the company of origin and/or the date of establishment of the company of arrival was missing. The table summarizes the number of job-to-job movements identified with geolocation (174) according to the quality of the combination (Table 2). For all of these cases, a manual check was then carried out.

Turning to the second criterion (firms' legal changes), in order to identify the existence of corporate transformations, we first looked at the number of individuals who made the same movement. Since we were observing a sample of about 5,000 workers, if there were, in the sample, at least three individuals with the same movement, it seemed reasonable to assume that it was spurious mobility. All these cases (88) therefore passed directly to manual check.

Table 2: Number of job-to-job according to the level of precision of the geolocalisation and the completeness of the available information

Geolocalisation quality	Cases with complete information	Cases with incomplete information [a]
High	42	69
Medium-high	7	33
Medium	2	1
Medium-low	0	3
Low	9	8
Total	174	

[a] Cases where the date of cessation of the company of origin and/or the date of establishment of the company of arrival were missing.

In cases where job-to-jobs occurred in the same month or on adjacent months (113 cases), a different procedure was followed based on variations in the size of the companies and, more precisely, on the monthly series of employees. The algorithm used for this purpose took into account the change in the number of employees at the time of the change of business: the delta of the company of origin (Δ company_from) corresponds to the difference in the number of employees between time t (where job-to-job occurs) and time t-1; likewise, the delta of the company of arrival (Δ company_at) corresponds to the difference in the number of employees between time t and time t+1.

The underlying logic is that the more similar the Δ were, the more likely it was that there were spurious mobility due to corporate transformations (such as spin off, company acquisitions, company branch acquisitions).

The cases identified (87) varied depending on the number of employees in the company of arrival (Nempl_at):

- |(Δ company _from – Δ company _at)|=< 2 if Nempl_at < 20
- |(Δ company _from – Δ company _at)|=< 5 if Nempl_at between 20 and 49
- |(Δ company _from – Δ company _at)|/ Nempl_at =< 0.1 if Nempl_at >= 50

Finally, cases of closure of the company of origin were isolated if the number of employees of the company of origin (Nempl _from) increased from N to 0 (6 cases):

- Δ company _from = Nempl _from

To sum up, the similarity of the attributes made it possible to identify 380 cases of probable spurious mobility, 206 on the basis of the name and 174 via

geolocation, respectively. On the other hand, 181 cases were identified from the analysis of workers' flows.

Considering that some cases were identified by more than one criterion, 476 company movements have been checked manually by consulting the websites of the companies involved (Table 3). Overall, 91.4% of the cases were positive, i.e. they confirmed the hypothesis of spurious mobility. The identification of 435 spurious company transfers made it possible to correct (and therefore not consider in the calculation of work mobility) 951 work episodes, corresponding to 3.8% of all the episodes in the private sector belonging to the regular fund.

Table 3: Number of possible spurious movements according to the criterion used to identify them and the result of the manual check.

Criterion	Manual check		N
	negative	positive	
Denomination	0	142	142
Geolocalisation	26	105	131
N. of employees	15	106	121
Denomination + Geolocalisation	0	22	22
Denomination + N. of employees	0	39	39
Geolocalisation + N. of employees	0	18	18
Denomination + Geolocalisation +N. of employees	0	3	3
Total	41	435	476

In conclusion, in order to create a database that would meet the purposes of analysing work histories, the procedures described above led to a reduction of about two thirds of the number of episodes that were originally in the administrative archive (from 158,000 to about 55,000).

4.3 THE MATCHING

As stated in Section 2 several content issues need to be considered while working with administrative and survey data (i.e. data definitions, consistency between different time periods, coverage in the administrative data system that may be subject to discontinuity due to changes in the legislation or in administrative practices).

In our case, the first important discrepancy concerned the *employment status* as the ICFT survey recoded all employment conditions, while INPS did not

include civil servants, freelancers with their own social security fund and, obviously, illegal workers (Figure 1).[15]

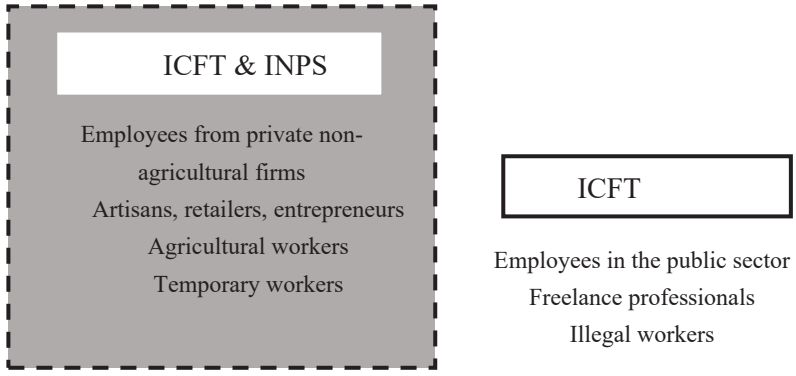


Figure 1: Type of occupations in the two sources of data

The second discrepancy concerned the *unemployment situation*. Determining unemployment episodes was problematic, for two reasons: the first one involved the difficulty to identify unemployment, as highlighted nearly fifty years ago by Shiskin (1976) in a paper with the paradigmatic title – [measuring] the Doughnut [of employment] or the Hole [of unemployment]? – and recurrent in the economic literature up to Brandolini and Viviano (2016). The second reason concerned the narrow perimeter of subsidized unemployment and the approximate way in which unemployment is recorded in the ICFT survey. Unlike in INPS, in the survey all those who declared themselves in search of a job were classified as unemployed (Figure 2). The survey therefore covered both those who were looking for their first job and those who were looking for a new job but without receiving any benefit.

The third factor of dissimilarity between the two sources concerned the *timing* of the observations: in INPS the episodes were documented with units of time at the day level (except for what has been said for the members of the separate fund), while in ICFT the segments of work histories declared by the interviewees were recorded with units of time by the month.

[15] Another possible underestimation is that of co.co.co and co.co.occ since for the former until January 1996 there was no obligation to register and for the latter it depends on the self-declaration of the worker.

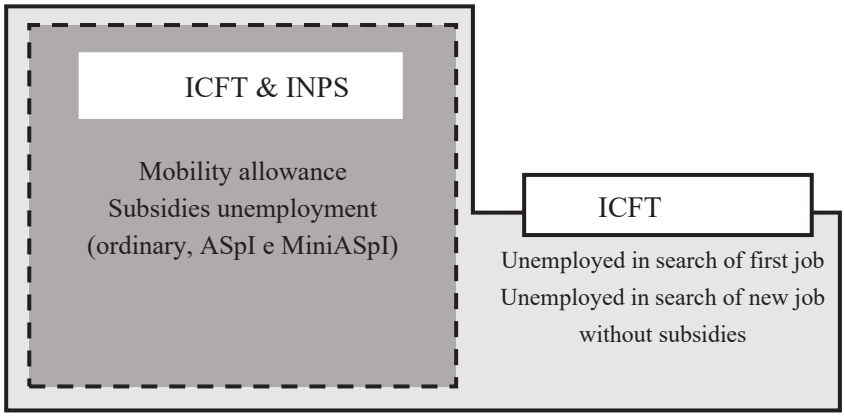


Figure 2: Non-working conditions in the two sources of data

The fourth and last source of distortion was linked to the fact that, while in the survey the prevailing condition was recoded excluding the *possibility of overlap* between episodes, in INPS it was possible to find (partial or complete) overlapping episodes.

In order to overcome the differences described above, when integrating the information taken from the two sources, a set of criteria was followed that privileged the employment condition and, overall, gave prevalence to the administrative source[16] (Figure 3).

As regards the working histories, the INPS source was treated as:

a. exclusive in the case of non-agricultural private employees.[17] It is worth mentioning that the number of episodes of this type detected in the survey was approximately one third of the total episodes.

[16] This in order to limit the possible response errors related to the survey.

[17] It should be noted that for 71 people who declared themselves to be non-agricultural private employees in the survey, no such episodes were found in INPS. In particular, 23 cases do not even appear in the administrative file, so their episodes of employment have been maintained but labelled as "illegal work". In the remaining cases, it was decided to eliminate only those episodes of non-agricultural private employment detected by interview if, in the same year, there was an episode of work in another category in INPS (thus assuming the existence of classification errors in the survey).

b. initial source, supplemented by ICFT, in the case of agricultural workers (self-employed and employees), craftsmen and traders and entrepreneurs (assuming that their absence was due to non-payment of contributions).

In the remaining cases, we resorted to:

c. ICFT data as an exclusive source for work episodes not to be found, by definition, in INPS (workers in public administrations and illegal workers);
d. Priority to ICFT for episodes of freelance professionals and temporary workers: (d1) the former because, if they had their own cash, they were present only in the ICVT; (d2) both because they could be found in INPS with significant information limits on the start and end dates.[18]

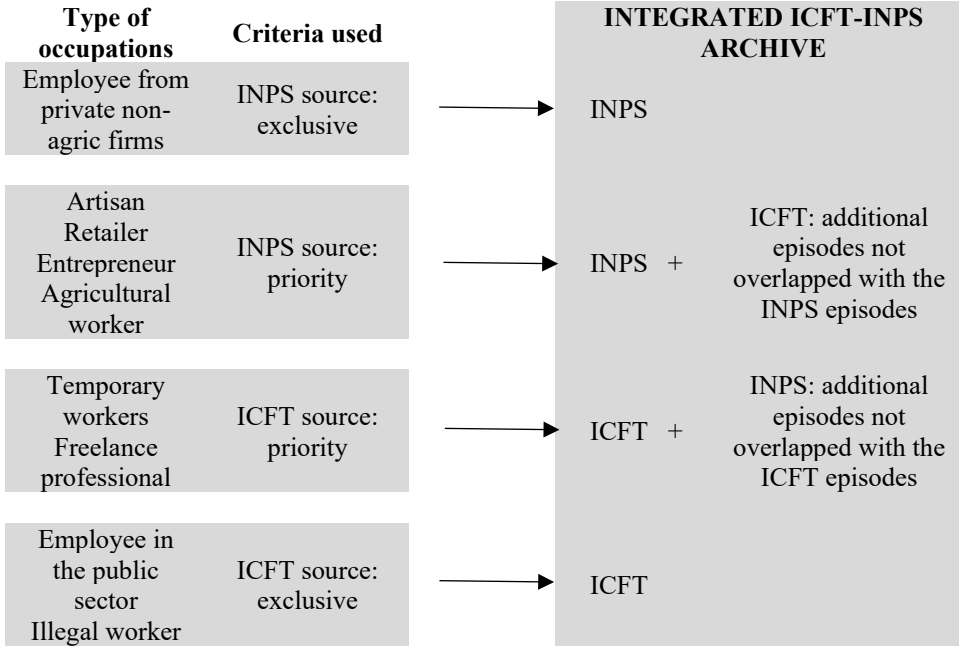


Figure 3: Data integration: exclusiveness or priority criteria used

[18] It should be noted that all the episodes belonging to the separate management, the beginning and the end, are conventionally equal respectively to 1 January and 31 December.

As for the unemployment episodes, the INPS priority criterion was used giving priority to the subsidised unemployment episodes. The integration with ICFT data was then used only if – or, in any case, for the part in which - the episode did not overlap with the periods of work as described above and with the periods of subsidized unemployment. In particular, two situations were identified. In cases (1,764) in which declared unemployment fell during a period in which neither work nor non-work episodes were found in INPS, a new episode of unemployment was added. In cases (193) where a period of declared unemployment was adjacent to a period of subsidised unemployment, the latter was extended to include the non-subsidised part.

Overall, the ICFT-INPS integrated archive comprised 5,489 individuals with at least one working episode. The working episodes were 48,580, of which 42,408 from the administrative archive, and 6,172 from the survey (Table 4). The unemployment episodes were, in total, 6,804:5,021 from INPS and the remaining from ICFT.

Table 4: Employment and unemployment episodes in the combined archive

Type	With no overlapping spells			Including overlapping spells		
	INPS	ICFT	Total	INPS	ICFT	Total
Employee from private non-agr. firms	29,648	0 [a]	29,648	34,959	0[a]	34,959
Employee in the public sector	n.d[b]	3,172	3,172	n.d[b]	3,367	3,367
Artisan, retailer and entrepreneur	1,379	188	1,567	1,465	204	1,669
Freelance professional	231	122	353	324	127	451
Temporary worker and other self-employed	1,743	249	1,992	3,205	263	3,468
Agricultural worker (employee & self-employer)	2,210	507	2,717	2,455	592	3,047
Illegal worker	n.d[b]	1,553	1,553	n.d[b]	1,619	1,619
Total employment spells	35,211	5,791	41,002	42,408	6,172	48,580
Total unemployment spells	5,000	1,783	6,783	5,021	1,783	6,804

[a] For the episodes of non-agricultural dependent work, a criterion of exclusivity of the administrative data was chosen, ignoring the episodes of that type declared in the survey.
[b] The episodes of public employment and "illegal" work were not included in INPS.

The table shows the number of episodes including overlapping spells and also those net of any overlaps between episodes of employment (double jobs) and between employment and unemployment. The criteria followed to solve overlapping spells were the following:

i) Where a working episode was completely contained in another more extensive working episode, the first was deleted.
ii) In cases of partial overlap between working episodes, a hierarchical order was first established between the types of work: private employee work (including agricultural work), public administration employee work, self-employment (including agricultural work), collaboration and undeclared work. On the basis of this, the starting or ending date of the lower hierarchical episode was changed.
iii) In the case of (non- part-time) employment overlapping with subsidised unemployment, unemployment had prevailed. The work episode was removed only if it was fully included in the unemployment episode. If, on the other hand, the unemployment episode was partially overlapped with a work episode, the latter was broken into two separate episodes. In case of partial overlap to the right or left of the work episode, the end or start date of the work episode was changed.
iv) In the case of overlap between part-time work and subsidised unemployment or mobility, work prevailed.[19]

Through the resolution of cases of overlap between episodes, the number of episodes was reduced by about 7,600 cases.

Regardless of whether or not the episodes overlap, it is worth mentioning that in addition to the private non-agricultural employees for whom the administrative data was considered an exclusive source, the contribution of the INPS data in the final integrated database was also dominant in cases where the survey data was considered a priority, and particularly in the case of collaborations and the liberal professions (Table 3).

The reasons may be twofold. First of all, as mentioned above, the survey recorded the prevailing working condition, therefore, in the case of double jobs (for example, an employee with a collaboration or the case of a teacher with

[19] The logic followed is the same as that described in the previous note with the only difference that, in this case, what is modified is the unemployment episode.

VAT) the respondents tended to declare the most important or the most stable over time.

The second reason is due to the fact that in the survey the work career was reconstructed retrospectively so the further away the event to be remembered, the more difficult it was for the respondent to identify short-term or occasional events.

The integration with the administrative data therefore had the undoubted advantage of correcting, at least for some types of episodes, the timing errors, completing the omissions of events and, in general, enriching the information base for the study of the dynamics of the working histories.

5. HIGHLIGHTS OF KEY FINDINGS

While the focus of this paper was to describe the construction and features of the data archive, it is worth briefly recalling some of the main findings derived from its analysis. These findings are explored in greater detail in a separate contribution, to which we refer for further discussion (Bazzoli et al. 2018).

The archive allowed to study changes in the work histories of young people over the last forty years. By comparing two cohorts - one composed of individuals born between 1959 and 1966, and the other of those born between 1975 and 1982 – we found an overall deterioration in working conditions for younger generations.

The average age of labour market entry increased from 18.2 years in the older cohort to 19.3 years in the younger one. This increase is particularly marked for women, who enter the labour market almost two years later than men in the most recent cohort. This delay is closely linked to longer education pathways, which have seen a marked improvement in educational attainment. In the younger cohort, women are on average more educated than men, reversing the trend observed in the previous generation.

Despite the improvement in education levels, the stability of first jobs has drastically decreased. The average duration of the first job declined from 23.6 months in the older cohort to 13.7 months in the younger cohort, with a particularly pronounced contraction among university graduates. The latter face the greatest difficulties in maintaining their first job, a phenomenon attributable to both their higher expectations in terms of job quality and pay, and the limited growth in demand for highly skilled labour. This reduction in first-job stability

affects both genders, although it is more pronounced for men, consistent with the changes in the sectoral distribution of first jobs.

Work histories of the younger cohort are also more fragmented and characterized by greater job mobility. During the first eight years of their careers, young people in the recent cohort experience a higher number of employment episodes compared to the older cohort, but with shorter durations. Pathways are particularly unstable for graduates and high school diploma holders, who face a growing number of fixed-term contracts and more limited access to stable positions. However, this increased mobility does not translate into higher career mobility, which remains low for both cohorts, highlighting persistent structural rigidity in the Italian labour market.

The comparison between the cohorts also reveals a polarization in types of employment. In the younger cohort, over 80% of young people begin their careers in the private sector, while the share of those entering public employment or self-employment has drastically declined. Women, compared to men, continue to have a higher likelihood of starting their careers in the public sector, although this tendency has diminished in the most recent generation. At the same time, the younger cohort faces a significantly higher risk of losing their first job, with a 57% increase compared to the older cohort. This risk is particularly high for more educated youth and those from families of dependent workers, whereas the children of self-employed workers demonstrate greater stability.

In summary, the comparison between the two cohorts highlights a deterioration in working conditions for younger generations. Despite higher levels of education, young people in the recent cohort face more unstable, fragmented, and precarious career paths, with increasing vulnerability in the labour market. This scenario reflects growing difficulties in aligning young people's qualifications with available opportunities, underscoring the structural challenges of a labour market characterized by polarization and inequalities.

6. CONCLUSIONS

As a result of the combination of restrictive interpretations of legislation already restrictive in itself and, even more so, the absence of an adequate scientific and political culture, the huge set of information contained in the archives of the Italian public administration has long remained unproductively hidden and only recently has begun to be used, albeit sporadically.

However, it should also be stressed that administrative archives alone are not always able to answer the research questions of scholars. Quite understandably these archives contain information about individuals that is segmented and filtered in the light of institutional needs. Moreover, they lack other significant data which, in many ways, are crucial in order to understand the configuration of what is contained in the administrative archive and, above all, to be able to use that data in order to carry out analyses and to evaluate public policies. To overcome the problem of segmentation of information, a way that has long been practiced in the Scandinavian countries already mentioned, but which is difficult to implement in Italy, is to combine complementary administrative archives. However, even this procedure, if it is truly feasible, is not always adequate to answer the research questions that arise from scientific research and the political sphere. An innovative and promising procedure for overcoming these possible further inadequacies is to integrate administrative archives with large-scale sample surveys using IT tools. This is the context in which the project on the working histories presented in this paper has started.

The integration between the INPS data and the ICFT panel survey as a tool for socio-economic analysis has confronted us with the multiple sources of potential distortion in the data of the two sources, the difficulties in identifying them, as a result of the myriad obstacles encountered in the integration at the micro scale. Often simplified solutions have been adopted (but plausible for the analysis of the changes also because it is reasonable to believe that the distortion is invariant - or exhibits low variability - over time). The analysis, however, has provided important insights into the deterioration of labour conditions for younger generations, the increasing mismatch between education levels and labour market opportunities, and the structural rigidity of the Italian labour market.

While challenges remain, this study represents a meaningful advancement in the integration of data for research and policy analysis in Italy. It contributes valuable methodological insights and demonstrates the potential of integrated datasets to improve our understanding of socio-economic dynamics.

REFERENCES

Al Baghal, T., Knies, G., Burton, J. (2014). Linking administrative records to surveys: differences in the correlates to consent decisions. *ISER, Understanding Society Working Paper Series* 9.

Bazzoli, M., Marzadro, S., Schizzerotto, A.,Trivellato, U. (2018). Come sono cambiate le storie lavorative dei giovani negli ultimi quarant'anni? Evidenze da uno studio pilota. *Stato e mercato*, 38(3), 369-418.

Brandolini, A., Viviano, E. (2016), Behind and beyond the (head count) unemployment rate. *Journal of the Royal Statistical Society*: Series A, 179(3): pp. 657-681.

Card, D., Chetty, R., Feldstein, M. S., Saez, E. (2010). *Expanding Access to Administrative Data for Research in the United States*. American Economic Association, Ten Years and Beyond: Economists Answer NSF's Call for Long-Term Research Agendas.

Calderwood, L., Lessof, C. (2009). Enhancing longitudinal surveys by linking to administrative data, in P. Lynn (Ed.), *Methodology of Longitudinal Surveys*, Chichester: Wiley: Ch. 4.

Connelly, R., Playford, C.J., Gayle, V., Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research,* 59, September: 1-12.

Conti, R., Bavaro, M., Boscolo, S., Fabrizi, E., Puccioni, C., Ricchi, O., Tedeschi, S. (2023). The Italian treasury dynamic microsimulation model (T-DYMM): Data, structure and baseline results. *Government of the Italian Republic (Italy), Ministry of Economy and Finance, Department of the Treasury Working Paper*, (1).

Crato, N., Paruolo, P. (2019). *Data-Driven Policy Impact Evaluation: How Access to Microdata Is Transforming Policy Design* (p. 346). Springer Nature.

Einav, L., Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1), 1-24.

Fabrizi, E., Rocca, A. (2024). NEET Status Duration and Socio-Economic Background. *Socio-Economic Planning Sciences*, 101986.

Fambri, G., Schizzerotto, A. (2008), a cura di. *Le condizioni di vita delle famiglie trentine. Rapporto di ricerca*, Quaderni della Programmazione n. 21, Provincia Autonoma di Trento e Università degli Studi di Trento, Trento: Edizioni 31.

Fondazione G. Brodolini, MEF "IESS Improving Effectiveness in Social Security Final Report" https://www.dt.mef.gov.it/export/sites/sitodt/modules/documenti_it/analisi_progammazione/analisi_programmazione_economico/final-report-IESS.pdf

Harron, K.L., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, L.M, Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, July–December: 1–12.

Istat (2009). Integrazione di dati campionari Eu-Silc con dati di fonte amministrativa. *Metodi e Norme* n. 38, Roma.

Künn S. (2015). *The Challenges of Linking Survey and Administrative Data*. IZA World of Labour, 214.

Livraga, G. (2019). Privacy in microdata release: Challenges, techniques, and approaches. *Data-Driven Policy Impact Evaluation*. Springer, 67-83.

Naticchioni, P., Raitano, M.,Vittori, C. (2016). La Meglio Gioventù: Earnings gaps across generations and skills in Italy. *Economia Politica*, *33*, 233-264.

Sakshaug, J. W., Couper, M. P., Ofstedal, M. B., Weir, D. R. (2012). Linking survey and administrative records: Mechanisms of consent. *Sociological Methods & Research* 41(4): 535–569.

Sestito, P., Trivellato, U. (2011). Indagini dirette e fonti amministrative: dall'alternativa all'ancora incompiuta integrazione. *Rivista di Politica Economica*, luglio-settembre: 183-227.

Shiskin, J. (1976), Employment and Unemployment: The Doughnut or the Hole?. *Monthly Labour Review*, 99(2): 3-10.

Trivellato, U. (2019). Microdata for social sciences and policy evaluation as a public good. In Crato N. and Paruolo P. (Eds).

doi.org/10.36253/ijas-16752

A MODIFIED CORRELATION BASED REGULARIZATION TECHNIQUE FOR REGRESSION ESTIMATION AND FEATURE SELECTION

Isaac Adeola Adeniyi, Dolapo Abidemi Shobanke

Department of Statistics, Federal University Lokoja, P.M.B. 1154, Lokoja, Nigeria

Abstract. *Variable selection is important for making sense with (ultra) high-dimensional data. Penalized least squares such as the LASSO, elastic-net and the correlation based elastic-net (L1CP) are popular methods for carrying out variable selection and estimation simultaneously. This study proposes a modified version of the L1CP motivated by reasons similar to that given by Zou and Hastie (2005) where the naïve elastic net was rescaled to give the elastic net. The scaling transformation is derived such that the double shrinkage caused by applying two penalties is undone thereby reducing bias. The derived scaling transformations are found to depend on the correlations among the predictors. A robust worst-case quadratic solver is used to obtain estimates. An evaluation of the proposed method which is referred to as CL1CP alongside the L1CP, LASSO and elastic-net through simulation studies illustrate the advantages of the CL1CP compared to the other alternatives considered especially in correct selection of sparse models. In terms of variable selection, estimation and prediction accuracy the proposed CL1CP performs favourably compared to the L1CP, LASSO and elastic-net especially for "grouped-variables" selection. Results from applications to two real life datasets corroborate the findings from simulation studies.*

Keywords*: Dimension reduction; Penalization; Sparsity; Data mining; Machine learning.*

1. INTRODUCTION

The multiple linear regression model is frequently used in studies for carrying out predictions and studying the effects of some predictor variables on a continuous response variable. Let the response vector be denoted by $Y = (y_1, \dots, y_n)^T$ and $X = [X_1, \dots, X_p]^T$ denote the predictor matrix, where p is the number of predictor variables. The regression model that represents the relationship between Y and $X_1, X_2, \cdots, X_p$ is

$$Y = \beta_0 + X^T\beta + \epsilon, \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown regression coefficients to be estimated with β_j representing the regression coefficient corresponding to predictor variable X_j, $j = 1, \ldots, p$; β_0 is the intercept and ϵ is a vector of error terms. Throughout this paper, we assume the errors ϵ are identically and independently distributed with zero mean and finite variance σ^2. The ordinary least squares (OLS) technique which is the classical approach for estimating the coefficients can be inefficient and not applicable when the predictors are highly correlated and when p is more than n respectively (Hoerl and Kennard, 1970; Ryan, 2009; Wang et al., 2016). With the advancement in technology, high and ultra-high dimensional data where number of variables p exceeds the sample size n are emerging from medical research and many other areas (Yahya et al., 2011; Hapfelmeier et al., 2012; Yahya et al., 2014; Yahya et al., 2015).

Variable selection is basically important for making sense with (ultra) high-dimensional data (Fan and Li, 2006) and it could lead to increased prediction performance of the fitted model. Classical model selection methods are the best-subset selection method and its step-wise variants. Nevertheless, best-subset selection is computationally impracticable when the number of predictors is large. Furthermore, as reported by Breiman (1996), subset selection is unstable while step-wise methods can perform very poorly resulting in a model with poor prediction accuracy. Also, when predictors are highly correlated, estimates of β by the OLS are not unique (Tan, 2012). The ridge regression (Hoerl and Kennard, 1970) uses an L_2-norm penalty to improve OLS when the predictors are correlated. The ridge estimator is obtained by solving the L_2-norm penalized least squares problem

$$\hat{\beta}_{ridge} = arg\min_{\beta} \|Y - \beta_0 - X^T\beta\|_2^2 + \lambda\|\beta\|_2^2$$

where $\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ is the L_2-norm of β. The ridge regression tries to take care of the problems caused by presence of correlated predictors in the model but does not carry out variable selection.

One of the most popular methods that have been proposed to overcome the underlying drawbacks of subset selection, stepwise selection and the OLS is the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996). The LASSO estimator is obtained by solving the L_1 penalized least squares problem

$$\hat{\beta}_{lasso} = arg\min_{\beta} \|Y - \beta_0 - X^T\beta\|_2^2 + \lambda\|\beta\|_1,$$

where $\|\beta\|_1^1 = \sum_{j=1}^{p} |\beta_j|$ is the L_1-norm of β. Although the LASSO has shown success in many situations, some of its limitations include its inability to select more than n predictors in the $p > n$ situation; tendency to select only one

variable from a group of predictors with high pairwise correlations and poor performance compared to ridge regression in the $n > p$ situation. To overcome the problems of LASSO highlighted above, Zou and Hastie (2005) proposed the elastic-net which is a combination of the L_1 penalty of the LASSO and the L_2 penalty of the ridge. The naïve elastic net estimator $\beta_{naive-enet}$ is the solution to

$$\hat{\beta}_{naive-enet} = \arg\min_{\beta} \|Y - \beta_0 - X^T\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2. \quad (2)$$

Zou and Hastie (2005) called the estimator obtained from (2) the naïve elastic-net estimator and showed that it does not perform satisfactorily nor possess the minimax optimal property because of the double shrinkage due to the L_1 and L_2 penalties which introduces unnecessary extra bias. Zou and Hastie (2005) obtained the elastic-net (ENET) estimator as $\left(1 + \frac{\lambda_2}{n}\right)\hat{\beta}_{naive-enet}$ and $(1 + \lambda_2)\hat{\beta}_{naive-enet}$ if the predictors are standardized (each variable has mean zero and l_2-norm one). However, Anbari and Mkhadri (2014) remarked that the ENET seems to be slightly less reliable if the correlation between variables is not so extreme. Besides, ENET ignores the information concerning the correlation of the data in the l_2 norm penalty.

With similar intent, the octagonal shrinkage and clustering algorithm for regression (OSCAR) which is based on the penalized least squares with a penalty function combining the L_1 and the pairwise L_∞ norms was proposed by Bondell and Reich (2008). The OSCAR constrains some coefficients to be identically equal, allowing correlated predictors that have similar effect on the response to form clusters represented by the same coefficients. Nevertheless, Anbari and Mkhadri (2014) observed that obtaining the OSCAR estimates can be slow for large p. Alternatively, Tutz and Ulbricht (2009) proposed a combination of a correlation based penalty (CP) and a blockwise boosting procedure (BB) for carrying out shrinkage and variable selection. But, in practice determining the step length factor and the stopping number of iterations for the blockwise boosting procedure can sometimes be tough and affects the sparsity of the solution as well as the speed of the procedure. As a result, Anbari and Mkhadri (2014) introduced the correlation based elastic-net estimator $\left(\hat{\beta}_{cp}\right)$ which is the minimizer of

$$\|Y - \beta_0 - X^T\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2 \sum_{j=1}^{p-1}\sum_{k>j}\left[\frac{\left(\beta_j - \beta_k\right)^2}{1 - r_{jk}} + \frac{\left(\beta_j + \beta_k\right)^2}{1 + r_{jk}}\right], \quad (3)$$

where r_{jk} is the empirical correlation between the predictors X_j and X_k. The L1CP is the combination of the LASSO and the correlation based penalty (CP) of Tutz and Ulbricht (2009). Similarly, Tan (2012) also proposed two correlation adjusted

elastic net (namely CAEN1 and CAEN2) penalties for linear regression and were further extended for the Poisson regression by Algamal and Lee (2015). The L1CP was found to have good empirical performance compared to the LASSO and ENET. Nonetheless, the double shrinkage incurred by using both the L_1 and correlation based penalties was not accounted for in the L1CP which can lead to unnecessary bias. Thus, the performance of the L1CP can be improved if the double shrinkage is undone.

Our goal in this study is to introduce scaled version of the L1CP. The main idea is to improve the performance of the L1CP by scaling the estimates in order to undo the double shrinkage obtained by combining both the L_1 and correlation based penalties. In this paper, we call the estimator proposed by Anbari and Mkhadri (2014) naïve L1CP while we call our proposed estimators CL1CP. We obtain the scales by carrying out a decomposition of the CP estimator (Tutz and Ulbricht, 2009) using ideas similar to those used by Zou and Hastie (2005) to correct the naïve ENET to give the ENET estimator.

The rest of the paper is organized as follows; In Section 2, we introduce the CL1CP and the estimators with corresponding proposed methods of rescaling. We carry out simulation studies to evaluate the finite sample performance of our proposals in comparison to the naïve L1CP as well as other competitors such as the LASSO and ENET in Section 3. In Section 4, we apply all the methods considered in Section 3 on real life data. Concluding remarks are given in Section 5.

2. THE METHODOLOGY- RESCALED L1CP (CL1CP)

In this section, we present the elements of the correction for the L1CP procedure to give the CL1CP. The objective function to be minimized to yield the L1CP regression estimates is

$$\|Y - \beta_0 - X^T\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2 P(\beta), \tag{4}$$

where

$$P(\beta) = \sum_{j=1}^{p-1}\sum_{k>j}\left[\frac{(\beta_j - \beta_k)^2}{1 - r_{jk}} + \frac{(\beta_j + \beta_k)^2}{1 + r_{jk}}\right].$$

Our proposed corrected estimator is of the form

$$\hat{\beta}_{CL1CP} = S \times \left[arg\min_{\beta}\|Y - \beta_0 - X^T\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2 P(\beta)\right], \tag{5}$$

where S is the scaling factor and r_{jk} is the empirical Pearson's product moment correlation between X_j and X_k. Details on the scaling factor are given in the

following. Tutz and Ulbricht (2009) showed that $P(\beta)$ can be written in a simple quadratic form:

$$P(\beta) = \beta^T W_{cp} \beta,$$

where $W_{cp} = \left(w_{ij}\right)_{1 \le i,j \le p}$ is a matrix such that

$$w_{ij} = \begin{cases} 2\sum_{s \neq i} \frac{1}{1 - r_{is}^2}, & i = j \\ \frac{-2r_{ij}}{1 - r_{ij}^2}, & i \neq j \end{cases},$$

and $r_{ij}^2 \neq 1, i, j = 1, \dots, p$. The estimator proposed by Anbari and Mkhadri (2014) which we call naïve L1CP estimator is a two-stage procedure: the correlation based penalty (CP) (Tutz and Ulbricht, 2009) regression coefficients are first obtained for each fixed λ_2, and then the lasso type shrinkage is carried out along the lasso coefficient solution paths. This leads to a double amount of shrinkage which introduces unnecessary extra bias, compared with pure CP or LASSO. We, therefore, introduce a corrected L1CP (CL1CP) estimator $\hat{\beta}_{CL1CP}$ which is given as

$$diag\left(\lambda_2 W_{cp} + I\right)\left[arg\min_{\beta} \|Y - \beta_0 - X^T\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2 \beta^T W_{cp} \beta\right], \qquad (6)$$

where I is a $p \times p$ identity matrix and $diag(A)$ is a diagonal matrix such that its diagonal elements are the same as the diagonal elements of A. It is easy to see that the CL1CP estimator is given by (5) when

$$S = diag\left(\lambda_2 W_{cp} + I\right). \qquad (7)$$

The scale factor for a single β_j can be simplified to

$$S_{jj} = 1 + 2\lambda_2 \sum_{s \neq j}^{p} \frac{1}{1 - r_{js}^2}.$$

The motivation for adopting the $diag\left(\lambda_2 W_{cp} + I\right)$ is similar to the argument used by Zou and Hastie (2005) to improve the performance of the ENET by rescaling the naïve ENET. First, it should be recalled that the correlation based penalty (CP) (Tutz and Ulbricht, 2009) estimator is given as

$$\hat{\beta}_{cp} = \left(X^T X + \lambda_2 W_{cp}\right)^{-1} X^T Y.$$

Now, let us consider a decomposition of the CP estimator with $\boldsymbol{X}$ standardized, we have

$$X^T X = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{12} & 1 & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \dots & 1 \end{bmatrix}.$$

Thus, the CP estimator can be rewritten as

$$\hat{\beta}_{cp} = R^{-1} \begin{bmatrix} 1 & r_{12}\left(1 - \frac{2\lambda_2}{1 - r_{12}^2}\right) S_1^{-1} & \dots & r_{1p}\left(1 - \frac{2\lambda_2}{1 - r_{1p}^2}\right) S_1^{-1} \\ r_{12}\left(1 - \frac{2\lambda_2}{1 - r_{12}^2}\right) S_2^{-1} & 1 & \dots & r_{2p}\left(1 - \frac{2\lambda_2}{1 - r_{2p}^2}\right) S_2^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p}\left(1 - \frac{2\lambda_2}{1 - r_{1p}^2}\right) S_p^{-1} & r_{2p}\left(1 - \frac{2\lambda_2}{1 - r_{2p}^2}\right) S_p^{-1} & \dots & 1 \end{bmatrix}^{-1} X^T Y, \quad (8)$$

where $R = \begin{bmatrix} S_1 & 0 & \dots & 0 \\ & S_2 & \dots & 0 \\ & & \ddots & \vdots \\ & & & S_p \end{bmatrix}$.

From (8), one can observe that L1CP estimator involves scaling the correlations r_{ij} by a factor $\left(1 - \frac{2\lambda_2}{1 - r_{ij}^2}\right) S_{jj}^{-1}$ followed by direct shrinkage with R^{-1} leading to double amount of shrinkage. When we combine the CP with the LASSO, the direct R^{-1} shrinkage step is not needed and removed by our scaling factor. In our proposed method, the shrinkage by LASSO is sufficient for controlling the variance and obtaining sparsity and we, therefore undo the R^{-1} shrinkage step by multiplying the naïve L1CP estimates by $diag(\lambda_2 W_{cp} + I)$. It can be easily shown that

$$R = \begin{bmatrix} S_{11} & 0 & \dots & 0 \\ & S_{22} & \dots & 0 \\ & & \ddots & \vdots \\ & & & S_{pp} \end{bmatrix} = diag(\lambda_2 W_{cp} + I).$$

2.1 Estimation and Selection of Tuning Parameters λ_1 and λ_2

The properties of the L1CP highlighted in Anbari and Mkhadri (2014) are preserved in our proposed scaling transformations. Anbari and Mkhadri (2014) showed that the L1CP can be augmented to become a LASSO type problem. Therefore, existing computational techniques for penalized regression methods can easily be used for obtaining the CL1CP. Following Anbari and Mkhadri (2014), the L1CP optimization problem can be written as

$$arg\min_{\beta^*}\left\|Y^* - \beta_0^* - X^{*T}\beta^*\right\|_2^2 + \lambda_1\|\beta^*\|_1, \tag{9}$$

where,

$X^*_{(n+p)\times p} = \begin{pmatrix} X \\ \sqrt{\lambda_2}L^T \end{pmatrix}, Y^*_{(n+p)} = \begin{pmatrix} Y \\ 0 \end{pmatrix}$ and $L^T = W_{cp}^{\frac{1}{2}}$.

L^T is obtained from the Cholesky decomposition of W_{cp}, that is, $LL^T = W_{cp}$. Note that W_{cp} is a real symmetric positive-definite square matrix, and that the L1CP estimator becomes the naïve Elastic-Net estimator when $W_{cp} = I$. See Anbari and Mkhadri (2014) for more details on the proof of how the L1CP method can be augmented to become LASSO type problems. The CL1CP estimator $\hat{\beta}_{CL1CP}$ can, therefore, be expressed as

$$diag(\lambda_2 W_{cp} + I) \times \left[arg\min_{\beta^*}\left\|Y^* - \beta_0^* - X^{*T}\beta^*\right\|_2^2 + \lambda_1\|\beta^*\|_1\right]. \tag{10}$$

It can be clearly seen that the CL1CP estimator becomes the Elastic-Net estimator when $W_{cp} = I$ whereas the L1CP can only reduce to the naïve Elastic-Net estimator. This is another justification for the need to rescale the L1CP as Zou and Hastie (2005) showed that the naïve Elastic-Net estimator does not perform adequately.

We adopt the quadratic solver method proposed by Grandvalet et al. (2017) to obtain the correlation based regression coefficients as well as the LASSO and elastic-net regression coefficients. In this approach, the inner minimization problem is seen as a simple unconstrained quadratic problem, and an optimization strategy based on the iterative resolution of plain quadratic problems is used to obtain minimizers of the objective function. The algorithm is highly computationally efficient and numerically robust compared to other optimization strategies for sparse regression such as coordinate descent (Fu, 1998) or proximal methods (Beck and Teboulle, 2009). The algorithm relies on second order techniques to solve a series of small-size quadratic approximations to the objective function defined from the currently estimated worst-case perturbation leading to stable convergence and reduced sensitivity to numerical errors, unlike first-order methods such as coordinate descent or proximal methods. The general

scheme of the algorithm starts from a vector of zeros and continues to keep an active set of variables. Only the active variables are involved when solving the underlying optimization problem, hence the size of the underlying problems solved is related to the number of nonzero coefficients in the vector of parameters. Therefore, by considering a decreasing grid of values for λ_1 and fixing λ_2, the entire path of solutions can be explored at a reasonable computational cost. The R (R Core Team, 2018) package *quadrupen* (Grandvalet et al., 2017) contains implementation of this algorithm. For more details on the algorithm, see Grandvalet et al. (2017).

Selecting the tuning parameters λ_1 and λ_2 is not a trivial task in practice and is very important in order to achieve a good prediction and estimation accuracy. The tuning parameters can be chosen by minimizing an estimate of the out-of-sample prediction error. A validation set can be used to estimate this directly if available, otherwise, one can use five-fold or ten-fold cross validation (Hastie *et al.*, 2001; Efron and Tibshirani, 1997; Tibshirani, 1996; Kohavi, 1995; Efron and Tibshirani, 1993). In this paper, we use ten-fold cross validation (10-fold CV) to select tuning parameters in applications on real life datasets while validation sets are generated to select tuning parameters in the simulation studies. The cross-validation (or validation) is done on a two dimensional surface since there are two tuning parameters in the ENET, L1CP and CL1CP. Usually a grid of λ_2 is chosen first, then for each λ_2 the entire solution path of the ENET, L1CP or CL1CP is produced by the quadratic solver algorithm. The second tuning parameter λ_1 is chosen by the 10-fold CV error. The chosen pair of λ_1 and λ_2 is the one with the smallest 10-fold CV error. See Hastie *et al.* (2001) for more information on selection of tuning parameters.

3. A SIMULATION STUDY

In this section, we present a simulation study to examine the performance of the CL1CP under various conditions with the OLS, stepwise, L1CP, LASSO and ENET. The methods are examined under different cases of medium, high and ultrahigh ($p > n$) dimensional settings. The true underlying regression model from which we simulate data is given by

$$Y = X^T\beta + \sigma\epsilon, \qquad \epsilon \sim N(0,1). \tag{11}$$

Each simulated data consists of a training set for fitting the model, a validation set for selecting the tuning parameters, and a test set on which the test errors are computed for evaluation of performance. The notation ·/·/· represents the number of observations in the training, validation, and test datasets, respectively. For

example, 100/200/300 indicates that the training, validation, and test datasets contain 100, 200, and 300 observations, respectively. In these simulation experiments, the smoothing parameters λ_1 and λ_2 were selected over a grid of 100 equally spaced values, ranging from 10^{-2} to $10^{2.2}$ for each parameter using the validation datasets.

3.1 Simulation Setting

The methods under consideration are examined under five simulation settings. The simulation settings represent different data scenarios and are similar to those used by Tutz and Ulbricht (2006) and Zou and Hastie (2005).

Setting 1: We simulated 100 data sets consisting of $n/10n/200$ observations and 8 predictors. We set $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)$, $n \in \{50, 100\}$ and $\sigma = 2$. The pairwise correlation between X_i and X_j was set to be $\rho(i,j) = \theta^{|i-j|}$ for all i, j, where $\theta \in \{0.5, 0.99\}$. This setting represents a sparse situation.

Setting 2: This setting is similar to that of Setting 1 except that $\beta_j = 0.85$, for all j.

Setting 3: In this instance, the simulated data sets consist of $n/10n/200$ observations and 40 predictors and we set $\beta = (\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10})$, $n \in \{50, 100\}$, $\sigma = 8$ and $\rho(i,j) = 0.5^{|i-j|}$ for all i, j. In this setting there are 40 sparse grouped predictors with only 20 being relevant.

Setting 4: In this setting, the simulated data sets consisting of $n/10n/200$ observations and 40 predictors and we set $\beta = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25})$, $n \in \{50, 100\}$ and $\sigma = 15$. The predictors X are generated as follows:

$X_i = Z_1 + w_i^x$, $Z_1 \sim N(0,1)$, $i = 1, \ldots, 5$,
$X_i = Z_2 + w_i^x$, $Z_2 \sim N(0,1)$, $i = 6, \ldots, 10$,
$X_i = Z_3 + w_i^x$, $Z_3 \sim N(0,1)$, $i = 11, \ldots, 15$.

X_i are independent identically distributed (iid) $N(0,1)$, for $i = 16, \ldots, 40$ and w_i^x are iid $N(0, 0.01)$. This setting implies there are three equally important groups with each containing 5 members.

Setting 5: In this situation, the simulated data sets consist of $n/10n/100$ observations and 200 predictors and we set $\beta = (\underbrace{5, \ldots, 5}_{20}, \underbrace{0, \ldots, 0}_{180})$, $n = 100$, $\sigma = 12$ and $\rho(i,j) = 0.5^{|i-j|}$ for all i, j.

3.2 Simulation Results

The performance of the methods are evaluated over 100 replications of each setting discussed above. The evaluation criteria are: prediction mean-squared

errors (MSE_Y) defined as $\frac{1}{n_{test}}\left\|Y_{test} - X_{test}^T\hat{\beta}\right\|^2$; mean-squared errors of estimates (MSE_β) defined as $\left\|\hat{\beta} - \beta\right\|^2$, size ($S$) which is the number of non-zero estimated regression coefficients; hits (C) which is the number of truly non-zero coefficients correctly estimated to be non-zero false positive (IC) which is the number of truly zero coefficients incorrectly estimated to be non-zero. For each method and simulation setting, each of the evaluation criteria was computed over 100 replications. Tables 1-5 summarizes the medians of MSE_Y, MSE_β, S, C and IC, while Figures 1-5 gives graphical representation of the results. The acronym TS stands for the true size of the model, that is, the number of important variables in the model.

The simulation results for Setting 1 (sparse situation) at $\theta = 0.5$ and $\theta = 0.99$ are presented in Table 1 and summarized in Figure 1 ($\theta = 0.5$, $n = 100$). The results show that OLS has the highest estimation and prediction error followed by Stepwise under all scenarios in Setting 1. When the correlations among the predictors are relatively low ($\theta = 0.5$), the CL1CP has the best performance in terms of prediction at both sample sizes considered. The performance of the ENET in terms of estimation at $n = 50$ is the best while that of CL1CP is the best at $n = 100$. The L1CP outperforms the OLS, Stepwise and LASSO in terms of prediction and estimation. In terms of variable selection, the CL1CP, ENET and Stepwise have the best performance while the L1CP and LASSO have similar performance. At high level of multicollinearity ($\theta = 0.99$) and $n = 50$, the L1CP and CL1CP have similar performance with respect to prediction but the L1CP seem to be slightly better than CL1CP in terms of estimation, while the LASSO have the worst performance in terms of prediction and estimation out of the penalized methods. Also, the CL1CP seem to be better than L1CP for variable selection at $\theta = 0.99$ and $n = 50$. Furthermore, at $\theta = 0.99$ and $n = 100$, the CL1CP is slightly better than L1CP in terms of prediction while for estimation L1CP seem to be slightly better. The LASSO has the best variable selection performance at $\theta = 0.99$ and $n = 50$ wth the L1CP and ENET having similar performances.

Table 1: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median absolute average errors of estimation (AE); median estimated model sizes (S), median Hits (C) and median FP (IC) for Setting 1 when $\theta = 0.5, 0.99$ based on 100 replications. TS stands for the true size of the model.

θ	n	*Method*	MSE_β	MSE_Y	S $(TS = 3)$	C	IC
0.5	50	***OLS***	0.88	4.77	-	-	-
		Stepwise	0.62	4.54	4	3	1
		LASSO	0.45	4.47	6	3	3
		ENET	0.28	4.26	4	3	1
		L1CP	0.44	4.43	6	3	3
		CL1CP	0.29	4.25	4	3	1
	100	***OLS***	0.43	4.33	-	-	-
		Stepwise	0.27	4.26	4	3	1
		LASSO	0.23	4.22	6	3	3
		ENET	0.14	4.14	4	3	1
		L1CP	0.22	4.22	6	3	3
		CL1CP	0.13	4.14	4	3	1
0.99	50	***OLS***	48.91	4.75	-	-	-
		Stepwise	35.44	4.59	3	1	1
		LASSO	9.65	4.19	4	2	2
		ENET	5.92	4.09	7	3	4
		L1CP	5.5	4.09	7	3	4
		CL1CP	5.6	4.09	6	3	3
	100	***OLS***	23.44	4.33	-	-	-
		Stepwise	19.89	4.31	3	2	1
		LASSO	6.38	4.16	4	3	2
		ENET	4.64	4.13	6	3	3.5
		L1CP	4.65	4.15	6	3	4
		CL1CP	4.71	4.11	7	3	4

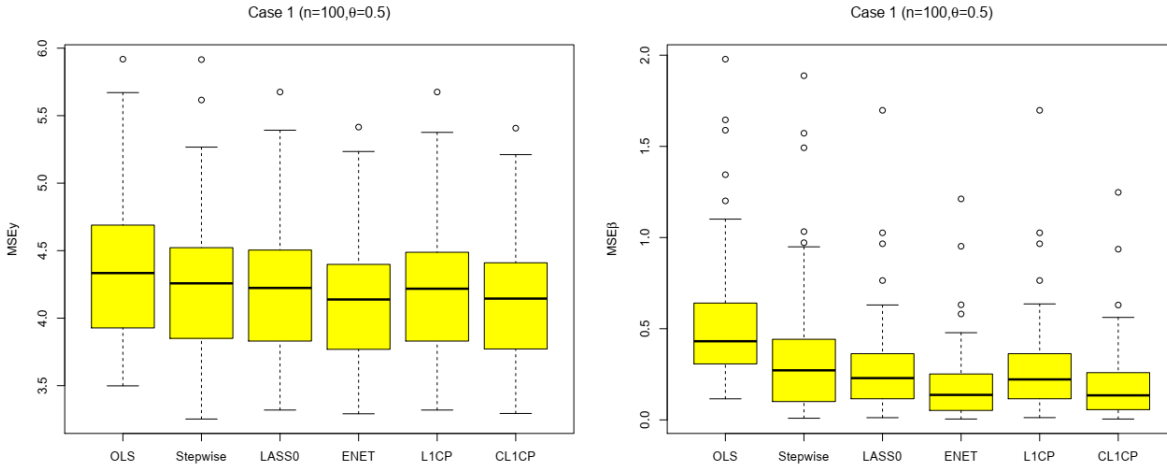


Figure 1: Boxplots of MSE_{β} and MSE_{Y} over 100 replications for Setting 1 at $n = 100$ and $\theta = 0.5$.

Table 2 and Figure 2 (for $\theta = 0.5$, $n = 100$) summarize the simulation results for Setting 2 (dense scenario). When the correlation among the predictors is relatively low ($\theta = 0.5$) and $n = 50$, LASSO has the worst performance in terms of estimation followed by OLS while the L1CP has the best performance in terms of estimation and prediction. The Stepwise has the worst performance at $n = 50$ in terms of prediction while CL1CP and the ENET have similar performances with respect to all the criteria in Setting 2. At $\theta = 0.5$ and $n = 100$, LASSO has the worst performance while in terms of variable selection, all the penalized methods perform equally. Furthermore, under Setting 2 when there is high correlation among the predictors ($\theta = 0.99$), the L1CP and CL1CP have similar performances in terms of variable selection at both sample sizes considered while Stepwise has poorest performance followed by the LASSO.

Table 2: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median absolute average errors of estimation (AE); median estimated model sizes (S), median Hits (C) and median FP (IC) for SETTING 2 when $\theta = 0.5, 0.99$ based on 100 replications.

θ	n	*Method*	MSE_β	MSE_Y	S $(TS = 3)$	C	IC
0.5	50	*OLS*	0.86	4.49	-	-	-
		Stepwise	1.39	4.77	7	7	0
		LASSO	0.88	4.46	8	8	0
		ENET	0.69	4.36	8	8	0
		L1CP	0.45	4.24	8	8	0
		CL1CP	0.69	4.34	8	8	0
	100	*OLS*	0.44	4.21	-	-	-
		Stepwise	0.45	4.26	8	8	0
		LASSO	0.43	4.30	8	8	0
		ENET	0.33	4.26	8	8	0
		L1CP	0.25	4.25	8	8	0
		CL1CP	0.34	4.26	8	8	0
0.99	50	*OLS*	44.89	4.49	-	-	-
		Stepwise	21.10	4.36	2	2	0
		LASSO	10.70	4.18	4	4	0
		ENET	0.09	3.93	5	5	0
		L1CP	0.07	3.96	8	8	0
		CL1CP	0.30	3.94	8	8	0
	100	*OLS*	23.58	4.21	-	-	-
		Stepwise	18.02	4.16	3	3	0
		LASSO	7.82	4.16	5	5	0
		ENET	0.05	4.07	8	8	0
		L1CP	0.03	4.03	8	8	0
		CL1CP	0.22	4.07	8	8	0

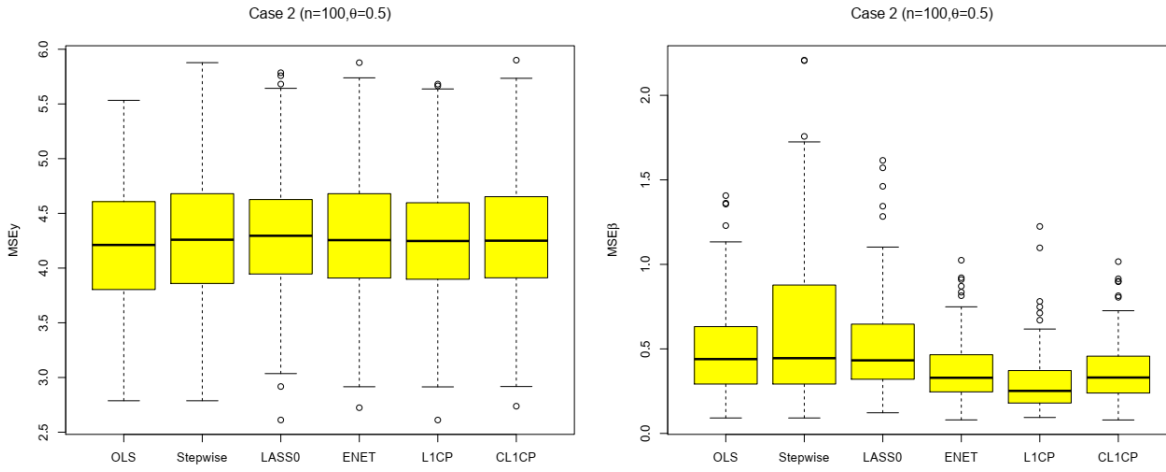


Figure 2: Boxplots of MSE_{β} and MSE_{Y} over 100 replications for Setting 2 at $n = 100$ and $\theta = 0.5$.

In terms of estimation and prediction, OLS and Stepwise perform very poorly while L1CP has the best performance in terms of estimation. At $n = 50$, the CL1CP is slightly better than L1CP in terms of prediction while at $n = 100$ the L1CP is the best. One can also observe that the performances of CL1CP and ENET are similar in terms of prediction but the LASSO is the poorest based on all criteria. It can also be noted that both the L1CP and CL1CP outperform the ENET in terms of variable selection at $n = 50$, an indication that having small sample size may affect the performance of the ENET and LASSO when the predictors are highly collinear. Under Setting 2, the L1CP mostly outperforms the CL1CP in terms of prediction and estimation, this is in line with the results of Zou and Hastie (2005) where the Naïve ENET outperforms the ENET under a similar setting like this.

From results for Settings 1 and 2, it appears that when there is high correlations among the predictors, the extra shrinkage incurred by the correlation based penalties is needed and should not be undone. Therefore, the naïve ENET and L1CP methods are better in terms of estimation and prediction when the correlations among the predictors are high and when all the predictors included in the model are relevant. However, the naïve methods are conservative and usually do not select sparse models compared to our proposed CL1CP as will see in the results from the remaining examples.

According to the Table 3 which presents the results for Setting 3 (equally grouped sparse variables) and Figure 3 (for $n = 100$), the CL1CP outperforms all the methods based on all the criteria with the ENET following closely and OLS and Stepwise having the worst performance. Of the penalized methods, the LASSO has the worst performance in terms of estimation and prediction when $n = 50$ while L1CP is the worst when $n = 100$. In terms of variable selection, the LASSO is better than the L1CP. The results here show that scaling the L1CP increases its performance in terms of estimation, prediction and ability to do "grouped variables" selection.

In Settings 4 and 5, the OLS and Stepwise methods are not included since they always behave poorly as in the previous settings and are not applicable in Setting 5 where $p > n$. The results for Setting 4 presented in Table 4 and summarized in Figure 4 ($n = 100$ only) reveal that L1CP outperforms all the other methods in terms of all the evaluation criteria at $n = 50$ followed by the ENET while the ENET has the best performance in terms of estimation and prediction at $n = 100$ followed closely by CL1CP. The LASSO performs poorly in terms of prediction, estimation and variable selection omitting relevant variables while the L1CP selects the highest number of irrelevant variables. The performance of the CL1CP here is significantly better than that of L1CP, this further shows that scaling the L1CP estimator improves its ability to do group selection.

The results for Setting 5 ($p > n$ situation) are summarized in Table 5 and Figure 5. According to the results, the ENET has the best performance in terms of estimation, prediction and variable selection followed by the LASSO and then CL1CP but in terms of variable selection. The results also show that the performance of the CL1CP is significantly better than that of the L1CP in terms of the criteria. The L1CP has poor performance in terms of estimation and prediction and selects a ridiculously large number of irrelevant variables compared to the CL1CP.

In general, the proposed CL1CP outperform the L1CP in almost all situations considered. The performance of the CL1CP does not strongly differ where the L1CP is better. The L1CP always select sparser models compared to the L1CP without omitting the relevant variables. Also, we find that the CL1CP outperforms the ENET and LASSO at times especially under situations where there are grouped variables and competes favourably in other situations.

Table 3: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median absolute average errors of estimation (AE); median estimated model sizes (S), median Hits (C) and median FP (IC) for SETTING 3 based on 100 replications.

n	*Method*	MSE_β	MSE_Y	S $(TS = 20)$	C	IC
50	***OLS***	5.29	7.36	-	-	-
	Stepwise	5.11	7.18	31	20	11
	LASSO	3.33	6.15	29	20	9
	ENET	2.41	5.47	23	20	3
	L1CP	3.09	6.11	31	20	11
	CL1CP	2.42	5.50	23	20	3
100	***OLS***	2.67	5.58	-	-	-
	Stepwise	2.26	5.34	25	20	5
	LASSO	1.58	5.17	28	20	8
	ENET	1.22	4.90	23	20	3
	L1CP	1.59	5.23	30	20	10
	CL1CP	1.21	4.90	23	20	3

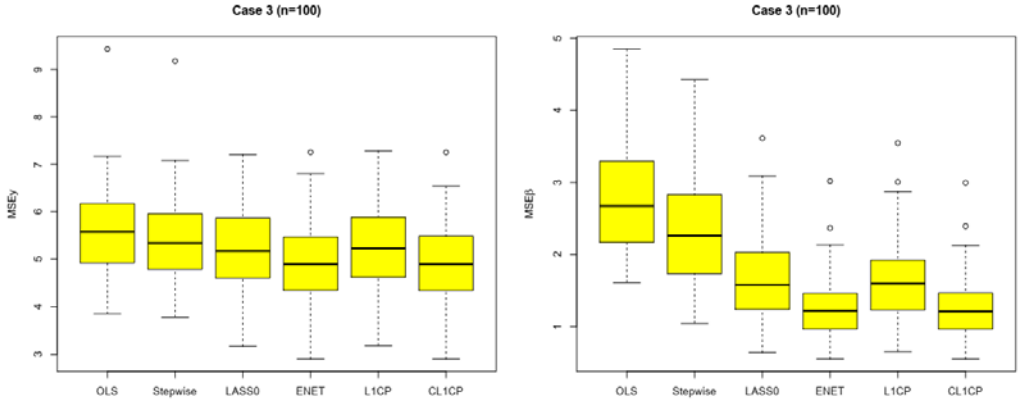


Figure 3: Boxplots of MSE_β and MSE_Y over 100 replications for Setting 3 at $n = 100$.

Table 4: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median absolute average errors of estimation (AE); median estimated model sizes (S), median Hits (C) and median FP (IC) for SETTING 4 based on 100 replications.

n	*Method*	MSE_β	MSE_Y	S $(TS = 15)$	C	IC
50	***LASSO***	385.67	288.97	10	3	7
	ENET	13.70	264.02	18	15	3
	L1CP	28.48	289.77	25	15	10
	CL1CP	11.95	258.58	18	15	3
100	***LASSO***	420.18	250.61	10	3	7
	ENET	2.74	235.40	17	15	2
	L1CP	9.95	249.53	23	15	8
	CL1CP	3.28	236.12	17	15	2

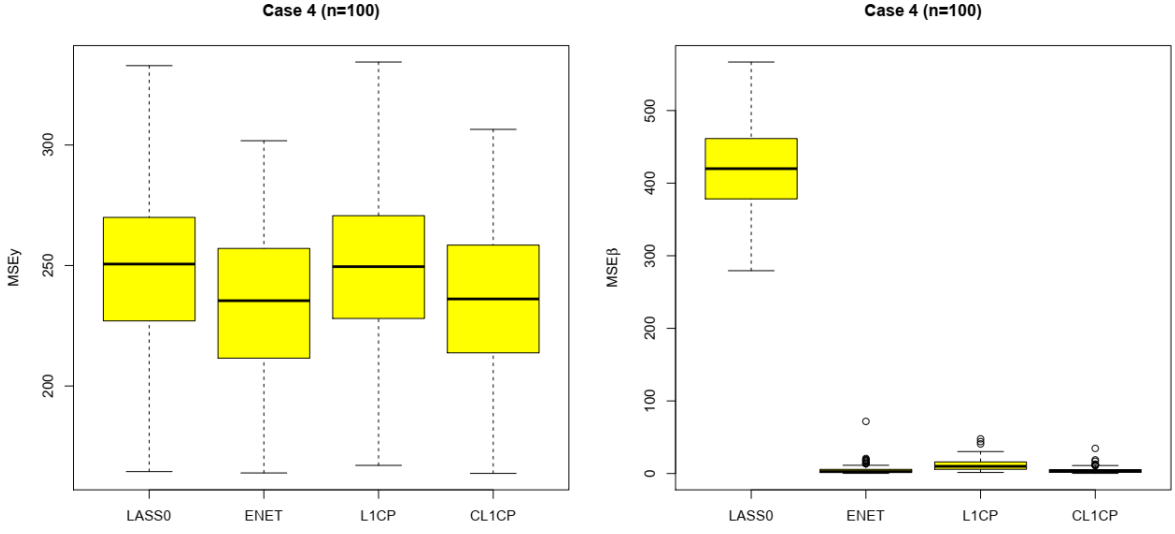


Figure 4: Boxplots of MSE_β and MSE_Y over 100 replications for Setting 4 at $n = 100$.

Table 5: Medians of mean squared errors of estimation and prediction (MSE_β and MSE_Y); median absolute average errors of estimation (AE); median estimated model sizes (S), median Hits (C) and median FP (IC) for SETTING 5 based on 100 replications at $n = 100$.

Method	MSE_β	MSE_Y	S $(TS = 15)$	C	IC
LASSO	105.04	246.57	43	20	23
ENET	68.94	216.65	42	20	22
L1CP	260.57	824.63	161	20	141
CL1CP	148.98	327.36	24.5	20	5

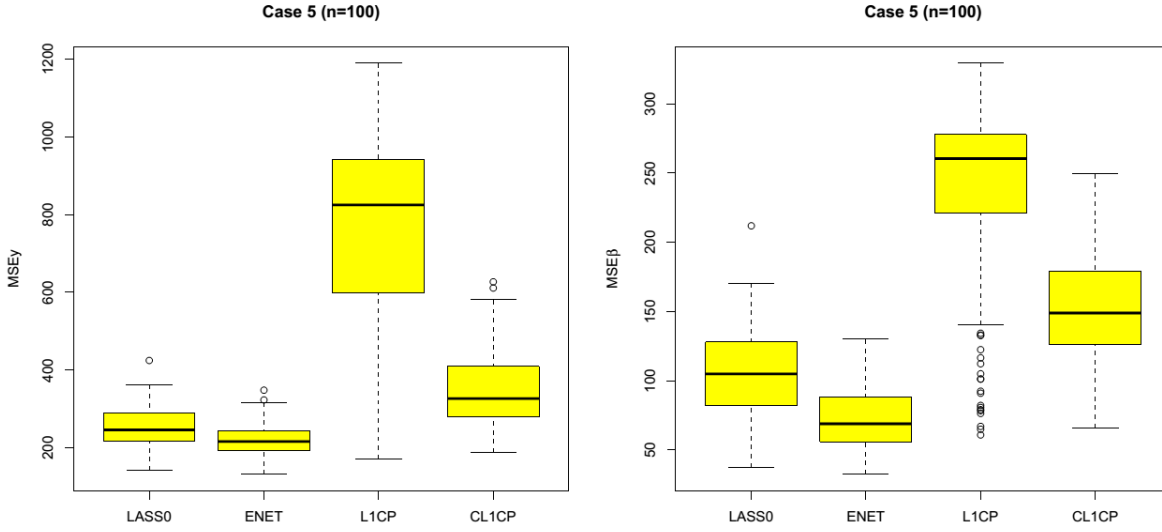


Figure 4: Boxplots of MSE_β and MSE_Y over 100 replications for Setting 5.

4. APPLICATION TO REAL LIFE DATASETS

In this section, applications of the methods on two real life datasets are considered. The first dataset is the prostate dataset used in Zou and Hastie (2005) which comes from a study of prostate cancer by Stamey et al. (1989) involving

97 men. The data consist of a response variable which is the log of prostate specific antigen (lpsa) and eight predictors: log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log capsular penetration (lcp), Gleason score (gleason) and percentage Gleason score 4 or 5 (pgg45).

The second dataset is the gene expression data from the microarray experiments on 120 mammalian eye tissue samples by Scheetz et al. (2006). The dataset consist of 200 predictors which represents 200 gene probes of 120 rats. The response variable is the expression level of TRIM32 gene. In contrast to the first dataset, the dimension of the gene expression data is very high with the sample size ($n = 120$) less than the number of predictors ($p = 200$). The OLS, Stepwise selection method, LASSO, ENET, L1CP and the CL1CP methods were all applied to the prostate dataset while only the penalized methods were applied to the gene expression dataset.

In the two applications presented in this section, the selection of the smoothing parameters λ_1 and λ_2 was performed using a grid search approach. Specifically, we considered a grid of 100 equally spaced values for each parameter, ranging from 10^{-2} to $10^{2.2}$. On a Windows 64-bit system with 8 GB RAM and an Intel Core i5-8365U CPU (1.60 GHz, 1.90 GHz turbo), the typical runtime for 10-fold cross-validation was 6 seconds for the prostate cancer dataset and 82 seconds for the eye tissue dataset. The Eye tissue dataset required more computational time than the prostate cancer dataset due to its higher dimensionality and larger sample size, which increased the complexity of the quadratic optimization and expanded the active set of variables during cross-validation, resulting in longer processing times. These results indicate that the validation process remains computationally efficient, making the proposed method practical for real-world applications, even in high-dimensional settings.

Firstly, the prostate cancer data were randomly split into a training set with 50 observations, and a test set with 47 observations. The training dataset were used for model fitting and selection of tuning parameters by10-foldcross validation. The performance of the methods were then compared based on their prediction mean squared error (MSEy) on the test dataset and the number of non-zero coefficients. For the gene expression dataset, the training set consists of 60 observations while the test set consists of 60 observations likewise. The process of data splitting, model fitting and computation of MSEy were repeated 100 times. The results for both datasets are summarized in Table 6.

Table 6 obviously shows that the non-corrected versions of the correlation based method (L1CP) selects denser models compared to the corresponding

corrected version (CL1CP) with no substantial gain in prediction accuracy in both datasets. The LASSO has the best performance in both datasets. For the prostate cancer data, the CL1CP selected the fewest number of predictors, which is same as the number of predictors selected by the stepwise method, and has prediction accuracy not significantly different from that of the L1CP which selects 3 variables more. Out of the penalized methods, the L1CP selects the highest number of predictors (176) with no significant gain in prediction accuracy.

In the case of the gene expression dataset, which is of ultrahigh dimension, the CL1CP outperforms all the methods in terms of sparsity with no significant loss in prediction accuracy. The L1CP produced a prediction error of about 0.007, albeit with median number of predictors of 145.5 which is almost six times that of the CL1CP (23.5 predictors) with prediction error of 0.008. So, the L1CP performs poorly in variable selection in cases of $p > n$ compared to CL1CP. Also, the CL1CP outperforms the ENET in terms of sparsity though they have the same prediction accuracy. The results from this section further show that the scaling the correlation based method (L1CP) improve its performances in terms of variable selection and can make it outperform the ENET.

Table 6: Median mean squared errors of prediction (MSE_Y) and median estimated model sizes (S), based on 100 replications.

Method	Prostate data		Eye tissue data	
	MSE_Y	S	MSE_Y	S
OLS	0.613	8	-	-
Stepwise	0.623	4	-	-
LASSO	0.594	5	0.008	23
ENET	0.597	5	0.008	31
L1CP	0.603	7	0.007	176
CL1CP	0.661	4	0.008	24

5. CONCLUDING REMARKS

In this paper, we have proposed a scaled version of the correlation based penalized elastic net penalty for carrying out variable selection and regression simultaneously. The motivation behind our idea is from Zou and Hastie (2005) where decomposition of the ridge operator was used to justify the need for

rescaling in other to ameliorate the challenges that are associated with double shrinkage through the use of the ridge and LASSO penalties. Following the ideas of Zou and Hastie (2005), we derive the scales by a decomposition of the correlation based estimators. We call the proposed scaled or corrected method CL1CP while the corresponding original version is referred to as the naïve L1CP. The efficient and robust worst-case quadratic solver method (Grandvalet et al., 2017) was adopted for estimation.

We have demonstrated via simulations and real life dataset that the proposed method compared to the naïve versions is better in terms of variable selection while retaining the ability to handle grouping effects and produce good prediction accuracy. The results also reveal that the scaled version of the correlation based method outperforms the LASSO and elastic-net in some situations especially in terms of variable selection in the presence of group effect. Generally, the new CL1CP proposed method possesses better performance in terms of correct identification of relevant variables which is vital in high dimensional regression problems. Finally, the new rescaled correlation based penalized regression method proposed in this work may attract wider applications in data mining and big data analytics.

A key advantage of the proposed methodology is its favourable computational efficiency. By leveraging an iterative quadratic solver that focuses only on the active set of variables, the algorithm significantly reduces the dimensionality of the optimization problem at each step. This selective approach results in a smaller computational cost compared to traditional sparse regression techniques such as coordinate descent or proximal methods.

The method proposed here may be extended to generalized linear models such as the binary logistic regression, Poisson regression, COM-Poisson regression (Shmueli et al., 2005; Adeniyi et al., 2019), and structured modelling (Ezenweke et al., 2023) settings by adopting a penalized likelihood approach.

Acknowledgments

We are grateful to two anonymous reviewers whose helpful reviews and suggestions improved the quality of this paper.

Source of Finance

During this study, no financial support was received neither from any individual or organization.

Conflict of Interest

There is no conflicts of interest declared by the authors.

REFERENCES

Adeniyi, I.A., Shobanke, D. A.,Edogbanya H. O. (2019). Re-parameterization of the COM-Poisson distribution using spectral algorithms. *Pakistan Journal of Statistics and Operation Research*, 15(3), 701-712.

Algamal, Z. Y., Lee, M. H. (2015). Penalized Poisson regression model using adaptive modified elastic net penalty. *Electronic Journal of Applied Statistical Analysis*, 8(2), 236-245.

Anbari, M.E., Mkhadri, A. (2014). Penalized regression combining the L_1 norm and a correlation based penalty**.** *Sankhya* B, 76 (1), 82–102.

Beck, A., Teboulle, M. (2009). Fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2,183–202.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statist.* 24, 2350–2383.

Bondell, H. D., Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR. *Biometrics* 64, 115 − 123.

Efron, B.,Tibshirani, R. (1993). *An Introduction to Bootstrap*. Chapman and Hall, London.

Efron, B., Tibshirani, R., (1997). Improvements on cross-validation. The 0.632+ bootstrap method. *Journal of the American Statistical Association*, 2, 548–560.

Ezenweke, C. P., Adeniyi, I. A., Yahya, W. B., Onoja, R. E. (2023). Determinants and spatial patterns of anaemia and haemoglobin concentration among pregnant women in Nigeria using structured additive regression models. *Spatial and Spatio-Temporal Epidemiology*, *45*, 100578. https://doi.org/10.1016/j.sste.2023.100578

Fan, J. Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *International Congress of Mathematicians*, **3,** 595–622.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3), 397–416.

Grandvalet, Y., Chiquet, J., Ambroise, C. (2017). Sparsity by worst-case quadratic penalties. *arXiv preprint arXiv:1210.2077v2 [stat.ML]*.

Hapfelmeier, A., Yahya, W.B., Rosenberg, R., Ulm, K. (2012). Predictive modelling of gene dxpression Data. In J. Crowley and A. Hoering (Eds), *Handbook of Statistics in Clinical Oncology*, 3rd ed. (pp. 463-475). Chapman and Hall/CRC, New York.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning*, Springer.

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

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection, *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1137–1145.

Ryan, T. (2009). *Modern Regression Methods (Second edition)*. John Wiley & Sons, Hoboken, NJ.

Scheetz, T. E., Kim, K. Y., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(39), 14429–14434.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., Boatwright, P. (2005). A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *Applied Statistics*, 54, 127–142.

Tan, Q. E. A. (2012). *Correlation Adjusted Penalization in Regression Analysis.* PhD. Thesis, The University of Manitoba, Canada. Retrieved from http://mspace.lib.umanitoba.ca/handle/1993/9147.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of*
the Royal statistical Society, Series B. 58, 267 − 288.

Tutz, g., Ulbricht, j. (2009). Penalized regression with correlation based penatly. *Statistics and. Computing.*, **19**, 239–253.

Wang, X., Dunson, D.B., Leng, C. (2016). No penalty no tears: Least squares in high-dimensional linear models. *Proceedings of Machine Learning Research,* 48, 1814-1822.

Yahya, W.B., Aremu, G.T. Garba, M.K. (2015). Multiclass sequential feature selection and classification method for genomic data. *Journal of Applied Science and Technology, 20* (1&2): 50-60.

Yahya, W.B., Rosenberg, R., Ulm, K. (2014). Microarray-based classification of histopathologic responses of locally advanced rectal carcinomas to neoadjuvant radiochemiotherapy treatment. *Türkiye Klinikleri Journal of Biostatistics, 6*(1), 8-23.

Yahya, W.B., Ulm, K., Fahrmeir, L., Hapfelmeier, A. (2011). k-SS: a Sequential feature selection and prediction method in microarray study. *International Journal of Artificial Intelligence*, 6(S11), 19-47.

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

doi.org/10.36253/ijas-16753

ON A WIDE CLASS OF GENERALIZED GEOMETRIC DISTRIBUTION FOR OVER AND UNDER DISPERSED DATA SETS

C. Satheesh Kumar[1]
Department of Statistics, University of Kerala, Thiruvananthapuram, India.

Harisankar, S
Asian School of Business, Technocity, Thiruvananthapuram, India.

Abstract *In this paper, a wide class of generalized geometric distribution is introduced and we name this class of distributions as "the alpha generalized geometric distribution (AGGD)". Several important distributions are obtained as a special cases of this proposed model. Important distributional properties such as generating functions, moments, recursive relations of the proposed distribution are examined. Parameter estimation using maximum likelihood is discussed. Three well-known data sets, having long tails, are analyzed and the results of fitting by various models are provided. Further, the generalized likelihood ratio test procedure is considered for testing the significance of the parameters of the GGD. Finally, performance of the different estimation methods are compared by means of a Monte Carlo simulation.*

Keywords: *Generalized likelihood ratio test; Model selection; Maximum likelihood estimation; Probability generating function; Simulation*

1. INTRODUCTION

Many generalizations of the geometric distribution have been attempted by researchers. Notable among them are Jain and Consul (1971) obtained a generalized negative binomial distribution. Philippou et al. (1983) developed a generalized geometric distribution and some of its properties. Tripathi et al. (1987) examined some generalizations of the geometric distribution derived by length biasing some generalized versions of the log-series distribution. Makcutek (2008) constructed a generalization of right truncated geometric distribution to modeling rank frequencies of graphemes. Gomez-Deniz (2010) proposed and studied a new generalization of the geometric distribution by using the Marshall and Olkin (1997) scheme.

[1] drcsatheeshkumar@gmail.com

Kumar and Harisankar (2020) considered a three parameter family of discrete distribution, namely "the doubly generalized Yule distribution (DGYD)" which we developed by compounding a geometric distribution with the generalized beta distribution, which combines two well known class of distributions namely GYD of Mishra (2009) and EYD of Martinez Rodriguez et al. (2011). We derived several important theoretical properties of DGYD which will help us to derive similar properties of both the classes of GYD and EYD. The p.g.f of the DGYD is the following.

$$G(t) = A_0 \, {}_2F_1(1, \eta; \gamma + \eta + 1; \lambda t), \tag{1}$$

where

$$A_0 = {}_2F_1(1, \eta, \gamma + \eta + 1; \lambda)^{-1}, \tag{2}$$

in which $\gamma > -(\eta + 1)$, $\eta > 0$ and $0 < \lambda < 1$.

Here,

$${}_2F_1(a, b, c; \lambda) = \sum_{r=0}^{\infty} \frac{(a)_r (b)_r}{(c)_r} \frac{\lambda^r}{r!},$$

is the Gaussian hypergeometric function (GHF); in which $(a)_r = a(a+1)\cdots(a+r-1)$ any positive integer r with $(a)_0 = 1$. For more details regarding GHF, see (Slater, 1966) or (Mathai and Haubold, 2008).

Kumar and Harisankar (2021) suggested another general class of distributions namely "the generalized geometric distribution (GGD)" which are more suitable for over-dispersed data sets, which obtained through compounding geometric and beta distribution and named it as "the generalised geometric distribution (GGD)". Further, we suggest an extended form of it and termed it as "the extended generalized geometric distribution (EGGD)" for creating more flexibility in modelling aspects. Here it can be seen that the standard geometric distribution is a special case of both the models. The p.g.f of the GGD is the following for $\gamma > -1$, $\eta > 0$ and $0 < \lambda < 1$.

$$H(t) = {}_2F_1[1, \eta; \gamma + \eta + 1; \lambda(t-1)], \tag{3}$$

Through this paper we propose a flexible count distribution as a generalization of both the DGYD and the GGD which we named as "the alpha generalized geometric distribution (AGGD)" and study some of its important properties. Since the proposed class of distribution enjoys both over-dispersed and under-dispersed

nature it can be viewed as a useful model in several practical situation. Its application in modeling empirical modeling of other count data is assessed by conducting comparative data fitting experiments. It is found that the proposed four parameter generalized geometric distribution performs well in all the examples considered here. In Section 2, we present the definition of the AGGD and derive its p.m.f., expression of generating functions, mean, variance, recursion formulae for its probabilities, raw moments and factorial moments. In Section 3, we discuss the estimation of the parameters of the AGGD by the method of maximum likelihood and consider generalized likelihood ratio test procedure for testing the significance of the parameters of the AGGD. So after investigating various useful properties, three data sets are considered here to justify its applicability. In Section 4, we carried out a simulation study for examining the performance of the maximum likelihood estimators of the parameters of the distribution.

Further we need the following series representation in the sequel.

$$\sum_{r=0}^{\infty}\sum_{s=0}^{\infty} A(s,r) = \sum_{r=0}^{\infty}\sum_{s=0}^{r} A(s,r-s). \tag{4}$$

2. PROPERTIES

Here, first we present the definition of the AGGD and discuss some of its important statistical properties.

Definition 2.1 *A non-negative integer valued random variable W is said to follow AGGD if its p.g.f is of the following form, in which* $\gamma > -(\eta+1)$, $\eta > 0$, $0 < \lambda < 1$, $-1 < \alpha < 1$ *and* $0 < \lambda + \alpha < 1$.

$$G_W(t) \quad = \quad A_0^* \, {}_2F_1(1,\eta;\gamma+\eta+1;\alpha+\lambda t), \tag{5}$$

where

$$A_i^* = [{}_2F_1(1+i,\eta+i,\gamma+\eta+1+i;\alpha+\lambda)]^{-1}, \quad for\ i = 0,1,2,... \tag{6}$$

The following table gives some of the special cases of AGGD.

Table 1: Special cases of the AGGD

No.	*Distribution*	*Parameter restriction*
1	Doubly generalized Yule distribution (DGYD) of (Kumar and Harisankar, 2020a)	$\alpha = 0$
2	Generalized geometric distribution(GGD) of (Kumar and Harisankar, 2020b)	$\alpha = -\lambda$
3	Hyper-logarithmic distribution (HLD) of (Tripathi and Gurland, 1985)	$\eta = 1$, $\alpha = 0$, $\gamma = \lambda - 1$
4	Generalized geometric distribution (GGD) of (Tripathi and Gupta, 1987)	$\eta = 2$, $\alpha = 0$, $\gamma = a - 2$
5	Extended Yule distribution(EYD) of (Martinez-Rodriguez et al., 2011)	$\eta = 1$, α=0
6	Generalized Yule distribution(GYD) of (Mishra, 2009)	α=0, $\lambda \longrightarrow 1$
7	Alpha logarithmic series distribution (ALSD) (Kumar and Riyaz, 2013a)	$\gamma = 0$, η=1
8	Alternative zero-inflated logarithmic series distribution (AZILSD) of (Kumar and Riyaz, 2013b)	$\alpha = -\lambda$, γ=0, $\eta = 1$
9	Yule distribution(YD) of Yule (1925)	$\eta = 1$, $\lambda \longrightarrow 1$, α=0
10	Geometric distribution	$\gamma = -1$, α=0
11	Logarithmic series distribution	$\gamma = 0$, η=1, α=0

Result 2.1 *The p.m.f* g_w *of AGGD with p.g.f* (5) *is the following, for* $w = 0,1,2,...,$ $\gamma > -(\eta+1)$, $\eta > 0$, $0 < \lambda < 1$ *and* $-1 < \alpha < 1$ *with* $0 < \lambda + \alpha < 1$.

$$g_w = A_0^* \ \frac{(\eta)_{w)} \ \lambda^w \ \beta_w(\alpha)}{(\gamma+\eta+1)_w}, \tag{7}$$

where $\beta_w(\alpha) = {}_2F_1(1+w, \eta+w, \gamma+\eta+1+w; \alpha)$ *and* A_0^* *is as defined in* (6).

Proof. From (5) we have the following:

$$G_W(t) = A_0^* \ {}_2F_1(1, \eta; \gamma+\eta+1; \lambda \ t+\alpha) \tag{8}$$

$$= \sum_{w=0}^{\infty} g_w \ t^w. \tag{9}$$

On expanding the Gauss hypergeometric function in (8), we get

$$G_W(t) = A_0^* \sum_{w=0}^{\infty} \frac{(\eta)_w}{(\gamma+\eta+1)_w} \ [\lambda \ t+\alpha]^w. \tag{10}$$

By applying (4) and binomial theorem in (10), we get the following.

$$\begin{aligned} G_W(t) &= A_0^* \sum_{w=0}^{\infty} \frac{(\eta)_w}{(\gamma+\eta+1)_w} \sum_{n=0}^{w} \binom{w}{n} (\lambda\ t)^{w-n}\ \alpha^n \\ &= A_0^* \sum_{w=0}^{\infty}\sum_{n=0}^{\infty} \frac{(\eta)_{w+n}}{(\gamma+\eta+1)_{w+n}} \binom{w+n}{n} (\lambda\ t)^w\ \alpha^n. \end{aligned} \tag{11}$$

On equating the coefficients of t^w on the right hand side expressions of (9) and (11) we get the following.

$$G_W(t) = A_0^* \frac{(\eta)_w \lambda^w}{(\gamma+\eta+1)_w} \sum_{n=0}^{\infty} \frac{(1+w)_n\ (\eta+w)_n\ \alpha^n}{(\gamma+\eta+1+w)_n\ n!}. \tag{12}$$

Hence the result. □

Result 2.2 *The moment generating function $M_W(t)$ of the AGGD is the following, for any $t \in R$*

$$M_W(t) = A_0^*\ {}_2F_1(1,\eta;\gamma+\eta+1;\alpha+\lambda\ e^t). \tag{13}$$

Result 2.3 *The cumulant generating function $K_W(t)$ of the AGGD is the following, for any $t \in R$ and $i = \sqrt{-1}$.*

$$K_W(t) = \ln[A_0^*] + \ln[{}_2F_1(1,\eta;\gamma+\eta+1;\alpha+\lambda\ e^{it})]. \tag{14}$$

Result 2.4 *The characteristic function $\psi_W(t)$ of the AGGD is the following, for any $t \in R$ and $i = \sqrt{-1}$.*

$$\psi_W(t) = A_0^*\ {}_2F_1(1,\eta;\gamma+\eta+1;\alpha+\lambda\ e^{it}). \tag{15}$$

Result 2.5 *The factorial moment generating function $F_W(t)$ of the AGGD with p.g.f is given by*

$$F_W(t) = A_0^*\ {}_2F_1[1,\eta;\gamma+\eta+1;\ \alpha+\lambda\ (t+1)]. \tag{16}$$

Result 2.6 *The mean and variance of the AGGD are the following,*

$$Mean = \frac{A_0^*\ \eta\ \lambda}{A_1^*(\gamma+\eta+1)} \tag{17}$$

and

$$Variance = \frac{A_0^*\ (\eta)\lambda}{A_1^*(\gamma+\eta+1)} \left[\frac{(\eta+1)\ \lambda}{A_2^*(\gamma+\eta+2)} + 1 - \frac{A_0^*\ \eta\ \lambda}{A_1^*(\gamma+\eta+1)}\right]. \tag{18}$$

Proof. It follows from the fact that,

$$Mean = G_W^1(1)$$

and

$$Variance = G_W^2(1) + G_W^1(1) - [G_W^1(1)]^2,$$

where $G_W^r(t) = \frac{d^r\, G(t)}{dt^r}/t = 1$. □

Result 2.7 *From (17) and (18), it can be seen that the AGGD is under-dispersed when $-(\eta+1) < \gamma < -\eta$ and over-dispersed when $\gamma > -\eta$ provided*

$$\frac{\eta\ (\eta+1)}{A_2^*(\gamma+\eta+2)} > \frac{A_0^*\,\eta}{A_1^*(\gamma+\eta+1)}\,.$$

Next we obtain certain recurrence relations for probabilities, raw moments and factorial moments of the AGGD. For the same we need the following notations, for $i = 0, 1, 2, ...$

$$\gamma^* + i = (1+i, \eta+i, \gamma+\eta+1+i) \tag{19}$$

$$R_i = \frac{(1+i)(\eta+i)}{(\gamma+\eta+1+i)}\,. \tag{20}$$

Result 2.8 *For $w \geq 1$, the following is a simple recurrence relation for probabilities $g_w = g_w(\gamma^*)$ of the AGGD with p.g.f (5).*

$$(w+1)\ g_{w+1}(\gamma^*) = \frac{A_0^*\, R_0\ \lambda}{A_1^*}\ g_w(\gamma^*+1), \tag{21}$$

in which R_0 and γ^* are defined in (19) and (20) respectively.

Proof. From (5), we have

$$G_W(t) = \sum_{w=0}^{\infty} g_w(\gamma^*)\ t^w = A_0^*\ {}_2F_1(1, \eta; \gamma+\eta+1; \lambda\ t+\alpha). \tag{22}$$

Differentiating the equation (22) with respect to t, we get

$$\sum_{w=0}^{\infty} (w+1)\ g_{w+1}(\gamma^*)\ t^w = A_0^*\, R_0\ \lambda\ \ {}_2F_1(2, \eta+1; \gamma+\eta+2; \alpha+\lambda\ t). \tag{23}$$

In (22) by replacing γ^* with γ^*+1 respectively, we obtain

$$\sum_{w=0}^{\infty} g_w(\gamma^*+1)\ t^w = A_1^*\ {}_2F_1(2, \eta+1; \gamma+\eta+2; \alpha+\lambda\ t). \tag{24}$$

Substitute (24) in (23) to get

$$\sum_{w=0}^{\infty}(w+1)\ g_{w+1}(\gamma^*)\ t^w = \frac{A_0^*\ R_0\ \lambda}{A_1^*}\ \sum_{w=0}^{\infty} g_w(\gamma^*+1)\ t^w. \tag{25}$$

Equating the coefficient of t^w on both sides of (25), we get (21). □

Result 2.9 *The following is a simple recurrence relation for raw moments* $\mu_r^{'} = \mu_r^{'}(\gamma^*)$ *of the AGGD, for* $r \geq 0$.

$$\mu_{r+1}^{'}(\gamma^*) = \frac{A_0^* R_0\ \lambda}{A_1^*} \sum_{s=0}^{r} \binom{r}{s}\ \mu_{r-s}^{'}(\gamma^*+1) \tag{26}$$

Proof. By definition, the characteristic function of the AGGD is given by

$$\psi_W(t) = \sum_{r=0}^{\infty} \mu_r^{'}(\gamma^*)\ \frac{(it)^r}{r!} \tag{27}$$

$$= A_0^*\ {}_2F_1(1,\eta;\gamma+\eta+1;\alpha+\lambda\ e^{it}). \tag{28}$$

By using (15) with γ^* replaced by γ^*+1 respectively, we obtain

$$\sum_{r=0}^{\infty} \mu_r^{'}(\gamma^*+1)\frac{(it)^r}{r!} = A_1^*\ {}_2F_1(2,\eta+1;\gamma+\eta+2;\alpha+\lambda\ e^{it}). \tag{29}$$

Differentiate (27) and (28) with respect to t, to get

$$\sum_{r=0}^{\infty} i\ \mu_{r+1}^{'}(\gamma^*)\frac{(it)^r}{r!} = A_0^*\ i\ R_0\ \lambda\ e^{it}\ {}_2F_1(2,\eta+1;\gamma+\eta+2;\alpha+\lambda\ e^{it}),$$

which on simplification gives

$$\sum_{r=0}^{\infty} \mu_{r+1}^{'}(\gamma^*)\frac{(it)^r}{r!} = \frac{A_0^*\ R_0\ \lambda\ e^{it}}{A_1^*} \sum_{r=0}^{\infty} \mu_r^{'}(\gamma^*+1)\frac{(it)^r}{r!}. \tag{30}$$

On expanding the exponential functions in (30) and applying (4) to obtain

$$\sum_{r=0}^{\infty} \mu_{r+1}^{'}(\gamma^*)\frac{(it)^r}{r!} = \frac{A_0^*\ R_0\ \lambda}{A_1^*} \left[\sum_{r=0}^{\infty}\sum_{s=0}^{r} \mu_{r-s}^{'}(\gamma^*+1)\frac{(it)^r}{(r-s)!s!}\right]. \tag{31}$$

Equating the coefficients of $(it)^r\ (r!)^{-1}$ on both sides of (31), we get (26). □

Result 2.10 *The following is a simple recurrence relation for factorial moments* $\mu_{[r]} = \mu_{[r]}(\gamma^*)$ *of the AGGD, for* $r \geq 0$.

$$\mu_{[r+1]}(\gamma^*) = \frac{A_0^* R_0 \lambda}{A_1^*} \mu_{[r]}(\gamma^* + 1). \tag{32}$$

Proof. The factorial moment generating function $F_{W_1}(t)$ of the AGYD is given by

$$\begin{aligned} F_W(t) &= \sum_{r=0}^{\infty} \mu_{[r]} \frac{t^r}{r!} && (33) \\ &= A_0^* \, {}_2F_1[1, \eta; \gamma + \eta + 1; \alpha + \lambda \ (t+1)] && (34) \end{aligned}$$

From (16) with γ^* changed by $\gamma^* + 1$ respectively, we have

$$\sum_{r=0}^{\infty} \mu_{[r]}(\gamma^* + 1) \frac{t^r}{r!} = A_1^* \, {}_2F_1[2, \eta + 1; \gamma + \eta + 2; \alpha + \lambda \ (t+1)]. \tag{35}$$

On differentiating (33) and (34) with respect to t, we get

$$\sum_{r=0}^{\infty} \mu_{[r+1]}(\gamma^*) \frac{t^r}{r!} = A_0^* R_0 \lambda \, {}_2F_1[2, \eta + 1; \gamma + \eta + 2; \alpha + \lambda \ (t+1)]. \tag{36}$$

Applying (35) in (36) to obtain

$$\sum_{r=0}^{\infty} \mu_{[r+1]}(\gamma^*) \frac{t^r}{r!} = \frac{A_0^* R_0 \lambda}{A_1^*} \left[\sum_{r=0}^{\infty} \mu_{[r]}(\gamma^* + 1) \frac{t^r}{r!} \right]. \tag{37}$$

By equating the coefficients of $t^r \ (r!)^{-1}$ on both sides of (37), we get (32). □

2.1. Estimation and Testing of Hypothesis

In this section we discuss the estimation of the parameters γ, η, λ and α of the AGGD by MLE and thereafter the GLRT procedure is utilized for testing the significance of the parameters γ, η, λ and α of the AGGD.

Let $a(w)$ be the observed frequency of w events based on the observations from a sample with independent components and let y be the highest value of the w observed. The likelihood function of the sample is

$$L = \prod_{w=0}^{y} [g_w]^{a(w)}, \tag{38}$$

which implies

$$\ln L = \sum_{w=0}^{y} a(w) \ln g_w. \tag{39}$$

Let $\hat{\gamma}$, $\hat{\eta}$, λ and $\hat{\alpha}$ be the MLEs of γ, η, λ and α respectively.Now the MLEs of the parameters are obtained by solving the following likelihood equations, obtained from (39) on differentiation with respect to γ, η, λ and α respectively and equating to zero. Then

$$\frac{\partial \ln L}{\partial \gamma} = 0$$

or equivalently

$$\begin{aligned}
&\sum_{w=0}^{y} a(w) \left[-A_0^* \sum_{r=0}^{\infty} \frac{(\eta)_r \, (\alpha+\lambda)^r}{(\gamma+\eta+1)_r} [\psi(\gamma+\eta+1) - \psi(\gamma+\eta+r+1)] \right. \\
&- \frac{1}{(\gamma+\eta+1)_w} (\psi(\gamma+\eta+1) - \psi(\gamma+\eta+w+1)) \\
&+ \frac{1}{\beta_w} \sum_{r=0}^{\infty} \frac{(1+w)_r \, (\eta+w)_r \, \alpha^r}{(\gamma+\eta+w+1)_r \, r!} \, [\psi(\gamma+\eta+w+1) \\
&\left. -\psi(\gamma+\eta+w+r+1)]] = 0
\end{aligned} \tag{40}$$

$$\frac{\partial \ln L}{\partial \eta} = 0$$

or equivalently

$$\begin{aligned}
&\sum_{w=0}^{y} a(w) \left[-A_0^* \sum_{r=0}^{\infty} \frac{(\eta)_r \, (\alpha+\lambda)^r}{(\gamma+\eta+1)_r} [\psi(\gamma+\eta+1) - \psi(\gamma+\eta+r+1) \right. \\
&+ \psi(\eta+r) - \psi(\eta)] + \frac{1}{(\eta)_w} [\psi(\eta+w) - \psi(\eta)] - \frac{1}{(\gamma+\eta+1)_w} [\psi(\gamma+\eta+w+1) \\
&- \psi(\gamma+\eta+w+r+1)] + \frac{1}{\beta_w} \sum_{r=0}^{\infty} \frac{(1+w)_r \, (\eta+w)_r \, (\alpha)^r}{(\gamma+\eta+w+1)_r \, r!} \\
&\left. \times [\psi(\gamma+\eta+w+1) - \psi(\gamma+\eta+w+r+1) + \psi(\eta+w+r) - \psi(\eta+w)]] = 0
\end{aligned} \tag{41}$$

$$\frac{\partial \ln L}{\partial \lambda} = 0$$

or equivalently

$$\sum_{w=0}^{y} a(w) \left[\frac{w}{\lambda} - \frac{A_0^* \, \eta}{(\gamma+\eta+1) \, A_1^*} \right] = 0 \tag{42}$$

and

$$\frac{\partial \ln L}{\partial \alpha} = 0$$

or equivalently

$$\sum_{w=0}^{y} a(w) \, [-\frac{A_0^* \, \eta}{A_1^* \, (\gamma+\eta+1)} - \frac{\beta_{w+1} \, (1+w) \, (\eta+w)}{\beta_w \, (\gamma+\eta+w+1)}] = 0, \tag{43}$$

where

$$\psi(\gamma) \quad = \quad [\Gamma(\gamma)]^{-1} \, \frac{d \, \Gamma(\gamma)}{d\gamma}.$$

It is important to note that the likelihood equations do not always guarantee a unique solution, and in some cases, the maximum of the likelihood function may occur at the boundary of the parameter domain. To address this, we computed the second-order partial derivatives of lnL with respect to the parameters α, λ, γ and η. Using MATHEMATICA, we verified that these second-order partial derivatives are consistently negative for the parameter ranges $\gamma > -(\eta+1)$, $\eta > 0$, $0 < \lambda < 1$ and $-1 < \alpha < 1$ with $0 < \lambda + \alpha < 1$. This confirms that the AGGD is log-concave within these restrictions, ensuring that the MLEs of α, λ, γ and η are unique and correspond to the global maximum of the likelihood function.

To further ensure that the MLEs do not occur at the boundary of the parameter space and that the regularity conditions for the validity of the asymptotic Chi-squared distribution are satisfied, we computed the Hessian matrix of lnL and verified that it is negative definite, further supporting the log-concavity of the likelihood function. Thus, the uniqueness and validity of the MLEs are established under the specified parameter restrictions. We have used the FindMaximum function of MATHEMATICA 11.2 then we get the maximum likelihood estimators of the parameters α, λ, γ and η of the AGGD.

Here we present the GLRT procedure for testing the significance of the parameters of the AGGD. Here we consider the following tests:

Test 1: ${H_0}^{(1)} : \alpha = 0$ against ${H_1}^{(1)} : \alpha \neq 0$

Test 2: ${H_0}^{(2)} : \alpha = 0, \eta = 1$ against ${H_1}^{(2)} : \alpha \neq 0, \eta \neq 1$

Test 3: ${H_0}^{(3)} : \alpha = 0, \lambda = 1$ against ${H_1}^{(3)} : \alpha \neq 0, \lambda \neq 1$

The test statistic is

$$-2 \ln \Lambda = 2 \left(\ln L(\underline{\hat{\Lambda}}; w) - \ln L(\underline{\hat{\Lambda}}^*; w) \right), \tag{44}$$

in which $\hat{\Lambda}$ is the MLE of $\underline{\Lambda} = (\alpha, \lambda, \gamma, \eta)$ with no restriction and $\hat{\underline{\Lambda}}^*$ is the MLE of $\underline{\Lambda}$ under H_0 . The test statistic (44) is asymptotically distributed as a χ^2 with one degree of freedom in case of Test 1 and two degrees of freedom in case of Test 2 and 3 respectively. We have computed the values of $\ln L(\hat{\underline{\Lambda}}; w)$, $\ln L(\hat{\underline{\Lambda}}^*; w)$ and the test statistic in case of all the three data sets and inserted in Table 5.

2.2. Applications

For numerical illustration, we have considered three real-life data applications, of which first data sets deals with the distribution of the plant Lespedeza Capitaita as reported by (Thomson, 1952) using contagious distributions. later it was studied by Katti and Gurland (1961). To illustrate the utility of the AGGD distribution we have considered a second data set about the number of public secondary schools in the 771 municipalities of Andalusia (Spain). The data is taken from Rodriguez et al. (2017). They have obtained the data from the database of the System of Multi territorial Information of Andalusia. Whereas, the third data represents the counts of cysts of kidneys using steroids from Chan et al. (2010). We have fitted the GGD (Kumar and Harisankar, 2020), the EYD (Martínez-Rodríguez et al, 2011), the AGGD and the Waring Distribution (WD) to these data sets and the results obtained along with the corresponding values of the expected frequencies, χ^2 statistic, d.f, AIC, BIC and AICc in respect of each of the models are presented in Tables (2), (3) and (4) respectively. Based on the computed values of the χ^2 statistic, AIC, BIC and AICc it can be observed that the AGGD gives better fit to all the three data sets considered here compared to the existing models the GGD, the EYD and the WD if the data is heavy-tailed.

Table 2: Observed frequencies and computed values of expected frequencies of the GGD, the EYD, the WD and the AGGD by MLE for the first data set.

w	Observed frequency	Expected frequency by MLE			
		GGD	EYD	WD	AGGD
0	7178	7156.29	6933.50	7128.28	7171.01
1	286	310	652.49	263.43	287.42
2	93	96.40	100.61	87.74	94.96
3	40	39.68	27.27	47.31	41.55
4	24	18.28	15.07	30.04	21.49
5	7	8.98	5.78	20.95	9.81

Continued ...

w	Observed frequency	Expected frequency by MLE			
		GGD	EYD	WD	AGGD
6	5	5.20	2.77	15.54	5.95
7	1	2.43	1.35	12.04	3.37
8	2	1.30	0.60	9.64	1.95
9	1	0.71	0.36	7.91	1.15
10	2	0.39	0.15	6.61	0.68
11	1	0.22	0.04	5.64	0.41
12	0	0.12	0.01	4.86	0.25
Total	7640	7640	7640	7640	7640
d.f		4	4	9	3
Estimates of parameters		γ=0.06 $\lambda = 0.62$ $\eta = 0.08$	γ=8.07 λ=0.99	γ=0.76 $\eta = 0.14$	γ=0.10 λ=0.65 $\eta = 0.14$ α=0.0005
χ^2 *value*		4.94	149.94	55.40	1.44
AIC		4602.80	4884.34	4608.70	4598.20
BIC		4599.98	4882.46	4606.82	4594.64
AICc		4604.80	4883.90	4609.80	4603.20

Table 3: Observed frequencies and computed values of expected frequencies of the GGD, the EYD, the WD and the AGGD by MLE for the second data set.

w	Observed frequency	Expected frequency by MLE			
		GGD	EYD	WD	AGGD
0	528	526.42	547.36	537.36	527.37
1	152	148.28	123.43	132.39	144.65
2	50	55.73	45.34	48.46	50.81
3	15	21.52	21.07	21.94	17.30
4	8	8.37	11.29	11.38	10.17
5	4	5.37	6.66	6.49	6.96
6	1	2.44	4.21	3.97	4.26
7	3	1.28	2.81	2.56	2.75
8	0	0.85	1.95	1.73	1.85

Continued ...

w	Observed frequency	Expected frequency by MLE			
		GGD	EYD	WD	AGGD
9	1	0.42	1.40	1.21	1.28
10	2	0.21	1.03	0.87	0.92
11	0	0.07	0.78	0.64	0.67
12	1	0.03	0.60	0.48	0.50
13	1	0.01	0.47	0.37	0.38
14	0	0.004	0.38	0.29	0.29
15	1	0.001	0.30	0.23	0.23
16	0	0.0007	0.25	0.18	0.18
17	0	0.0003	0.20	0.15	0.14
18	0	0.0001	0.17	0.12	0.12
19	1	0.00004	0.14	0.10	0.09
20	3	0.0000	0.12	0.08	0.08
Total	771	771	771	771	771
d.f		3	5	5	3
Estimates of parameters		γ=-0.91 $\lambda = 0.40$ $\eta = 0.22$	γ=2.39 λ=0.99	γ=3.0 $\eta = 1.31$	γ=2.82 λ=0.96 $\eta = 1.38$ α=0.0065
χ^2 *value*		17.25	11.59	9.64	2.11
AIC		1718.64	1619.08	1622.80	1616.92
BIC		1716.40	1617.72	1621.44	1614.20
AICc		1719.84	1619.48	1623.40	1619.42

Table 4: Observed frequencies and computed values of expected frequencies of the GGD, the EYD, the WD and the AGGD by MLE for the third data set.

w	Observed frequency	Expected frequency by MLE			
		GGD	EYD	WD	AGGD
0	65	57.49	61.40	63.61	65.19
1	14	19.27	20.17	20.95	14.27
2	10	12.62	10.05	9.68	9.55
3	6	8.84	5.76	5.77	6.44

Continued ...

w	Observed frequency	Expected frequency by MLE			
		GGD	EYD	WD	AGGD
4	4	5.12	4.23	3.21	4.23
5	2	3.08	3.11	2.10	3.11
6	2	1.76	1.67	1.45	2.49
7	2	0.94	1.20	1.04	1.64
8	1	0.48	0.88	0.77	1.28
9	1	0.25	0.65	0.59	0.85
10	1	0.11	0.50	0.46	0.51
11	2	0.04	0.38	0.37	0.44
Total	110	110	110	110	110
d.f		2	3	2	1
Estimates of parameters		γ=-0.54 $\lambda = 0.63$ $\eta = 2.62$	γ=0.60 λ=0.88	γ=1.91 $\eta = 1.50$	γ=-0.86 λ=0.73 $\eta = 0.15$ α=0.001
χ^2 *value*		4.85	2.55	4.86	0.52
AIC		358.88	346.84	348	342.06
BIC		356.18	345.04	346.20	338.46
AICc		360.38	348.14	349.30	347.86

Table 5: Test statistic value and table value for GLRT in case of all the three data sets

Data sets		$\ln L(\hat{\underline{\Lambda}}; w)$	$\ln L(\hat{\underline{\Lambda}}^*; w)$	Test Statistic	d.f	Tabled value
Data set 1	Test 1	-2292.10	-2295.40	6.60	1	3.84
	Test 2	-2295.10	-2440.17	290.14	2	5.99
	Test 3	-2295.10	-2302.35	14.50	2	5.99
Data set 2	Test 1	-805.46	-856.22	101.52	1	3.84
	Test 2	-805.46	-809.54	8.16	2	5.99
	Test 3	-805.46	-809.40	9.80	2	5.99
Data set 3	Test 1	-167.03	-176.44	18.82	1	3.84
	Test 2	-167.03	-171.42	8.78	2	5.99
	Test 3	-167.03	-172	9.94	2	5.99

From Table 5, it is observed that, for all three tests, the calculated value of the

test statistic exceeds the critical value corresponding to the 0.95 quantile of the Chi-squared distribution. Therefore, it can be concluded that the parameters of the fitted AGGD model are significant across all three datasets at the 5% level of significance.

2.3. Simulation

To examine the performance of the MLEs, a simulation procedure was conducted for different sample sizes ($n = 100, 200, 500$). We simulated 1000 samples from the AGGD, EYD and WD and then estimated the parameters by the maximum-likelihood method. The true parameter values used in the data generating processes for AGGD are i) γ= -0.45, $\eta = 0.55$, α=0.25, λ=0.05 (under-dispersed) ii) γ= 0.30, $\eta = 1.20$, α=0.45, λ=0.02 (over-dispersed). The true parameter values utilized in the data generation processes for EYD are i) γ= -0.25, λ=0.08 (under-dispersed) ii) γ= 0.35, λ=0.12 (over-dispersed) and The true parameter values applied in the data generation processes for WD are i) γ= 0.45, $\eta = 0.55$ (over-dispersed) ii) γ= 0.30, $\eta = 1.20$ (over-dispersed). By using simulated observations, we estimated the parameters of the AGGD, the EYD and the WD and thereby computed the values of the absolute bias and mean squared errors of each of the estimators. The results obtained are presented in Tables 6, 8 and 7 from which it can be observed that both the absolute values of bias and MSE of the estimators of the parameters are in decreasing order as the sample size increases. The simulation results demonstrate that the AGGD outperforms the other two distributions in terms of bias and mean square error (MSE), particularly as sample size increases. The lower bias indicates that the AGGD provides more accurate estimates closer to the true parameter value, while the reduced MSE reflects improved precision and accuracy, combining lower variability with minimal systematic error. This suggests that AGGD exhibits strong asymptotic properties, making it highly effective for larger datasets. Its superior performance highlights its potential as a robust and reliable method for parameter estimation, contributing to statistical methodology and practical applications where accuracy and precision are critical.

Table 6: Absolute bias and MSE (in the parenthesis) of the estimators of the parameters γ, η, α and λ of the AGGD for the simulated data sets.

		MLE			
Parameter set	*Sample size*	$\hat{\gamma}$	$\hat{\eta}$	$\hat{\alpha}$	$\hat{\lambda}$
(i)	$n = 100$	0.31 (0.262)	0.40 (0.346)	0.12 (0.081)	0.06 (0.065)
	$n = 200$	0.18 (0.185)	0.24 (0.142)	0.06 (0.041)	0.01 (0.024)
	$n = 500$	0.09 (0.064)	0.16 (0.041)	0.004 (0.0012)	0.005 (0.0011)
(ii)	$n = 100$	0.17 (0.111)	0.61 (0.282)	0.15 (0.051)	0.036 (0.0013)
	$n = 200$	0.08 (0.082)	0.31 (0.104)	0.04 (0.0021)	0.0051 (0.0002)
	$n = 500$	0.03 (0.0001)	0.11 (0.024)	0.008 (0.0006)	0.0002 (0.00008)

Table 7: Absolute bias and MSE (in the parenthesis) of the estimators of the parameters γ and λ of the EYD for the simulated data sets.

		MLE	
Parameter set	*Sample size*	$\hat{\gamma}$	$\hat{\lambda}$
(i)	$n = 100$	0.45 (0.462)	0.48 (0.46)
	$n = 200$	0.38 (0.285)	0.44 (0.242)
	$n = 500$	0.21 (0.094)	0.28 (0.082)
(ii)	$n = 100$	0.55 (0.25)	0.92 (0.42)
	$n = 200$	0.44 (0.182)	0.31 (0.164)
	$n = 500$	0.24 (0.08)	0.20 (0.12)

Table 8: Absolute bias and MSE (in the parenthesis) of the estimators of the parameters γ and η of the WD for the simulated data sets.

		MLE	
Parameter set	*Sample size*	$\hat{\gamma}$	$\hat{\eta}$
(i)	$n = 100$	0.65 (0.562)	0.98 (0.66)
	$n = 200$	0.52 (0.451)	0.66 (0.352)
	$n = 500$	0.30 (0.221)	0.42 (0.149)
(ii)	$n = 100$	0.45 (0.28)	0.72 (0.38)
	$n = 200$	0.38 (0.176)	0.56 (0.155)
	$n = 500$	0.23 (0.10)	0.44 (0.13)

3. SUMMARY and CONCLUSION

In this paper, we introduced the alpha generalized geometric distribution (AGGD), a univariate distribution that demonstrates remarkable flexibility in modeling both over-dispersed and under-dispersed datasets. The AGGD encompasses several important generalized geometric models and well-known generalized distribu-tions as special cases, highlighting its versatility.

We derived key statistical properties of the distribution and estimated its parameters using the method of maximum likelihood estimation (MLE). A test procedure was proposed to evaluate the significance of the parameters, and the results were compared with those obtained from existing models, showcasing the advantages of the AGGD.

To validate the performance of the proposed estimators, a simulation study was conducted, confirming their effectiveness under different scenarios. The findings suggest that the AGGD provides a compelling alternative to existing models for the types of datasets considered in this study.

While this paper presents several characteristic properties and inferential aspects of the model, further exploration of its properties and applications remains an avenue for future research, which we aim to publish in due course.

REFERENCES

Chan, S. K., Riley, P. R., Price, K. L., McElduff, F., Winyard, P. J., Welham, S. J., Woolf, A. S., Long, D. A. (2010). Corticosteroid-induced kidney dysmorphogenesis is associated with deregulated expression of known cystogenic molecules, as well as Indian hedgehog. *American Journal of Physiology-Renal Physiology*, 298(2), 346-356.

Gomez-Deniz, E. (2010). Another generalization of geometric distribution. *Test*, 19, 399-415

Gupta, R. C., Gupta, P. L., Gupta, R. D. (1998). Modeling failure time data by Lehman alternatives. *Commun. Statist. Theor Meth*, 39(15), 2767-2788.

Gupta, R. D., Kundu, D. (2007). Generalized exponential distribution: Existing results and some recent developments. *J. Statist. Plann. Infer*, 137, 3537-3547.

Jain G. C., Consul, PC (1971) A generalized negative binomial distribution. *SIAM J Appl Math*, 21, 501-513.

Katti, S.K., Gurland, J. (1961). The Poisson Pascal distribution. *Biometrics,* 17(4), 527-538.

Kumar, C.S., Riyaz, A. (2013). An extended zero-inflated logarithmic series distribution and its application, *Journal of Applied Statistical Science*, 21(1), 31-42.

Kumar, C.S., Riyaz, A. (2013). On Zero-Inflated logarithmic series distribution and its application, Statistica , 73(4), 477-492.

Kumar, C.S. and Riyaz, A. (2013). An alternative to the logarithmic-with- zeros distribution and its properties and estimation. *South African Statistical Journal*, 48(2), 253-266.

Kumar, C.S., Harisankar. S. (2020). On some aspects of a general class of Yule distribution and its Applications, Commun. Statist. Theor Meth, 49(12), 2887-2897.

Kumar, C.S. and Harisankar. S. (2020). Modified geometric distribution: properties and applications , *Stochastic Modelling and Applications*, 24(1), 4

Kumar, C. S., Harisankar, S. (2020). Modeling over-dispersed datasets using generalized geometric distributions. *Journal of Statistical Computation and Simulation*, 90(4), 606-623.

Makcutek J (2008). A generalization of the geometric distribution and its application in quantitative linguistics. *Romanian Rep Phys*, 60(3), 501-509.

Marshall, A.W, Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull fam- ilies. *Biometrika,* 84(3), 641-652.

Martínez-Rodríguez A.M., Sáez-Castillo, A. J., Conde-Sánchez, A. (2011). Modelling using an extended Yule distribution. *Computational Statistics and Data Analysis,* 55(1), 863-873.

Mathai, A. M., Haubold, H. J. (2008). Special Functions for Applied Statistics. *Springer*, New York .

Mishra, A. (2009). On a generalized Yule distribution. *Assam Statistical Re- view,* 23, 140-150.

Mudholkar, G. S, Srivastava, D. K. (1993). Exponentiated Weibull family for analyzing bathtub failure data. *IEEE Trans. Relia*b, 42, 299-302.

Philippou, A.N, Georghiou, C., Philippou, G.N. (1983). A generalized geo- metric distribution and some of its properties. *Stat Probab Lett*, 1, 171-175.

RodrÃguez-Avi, J., Olmo-JimÃ©nez, M. J. (2017). A regression model for overdispersed data without too many zeros. *Statistical Papers*, 58(3), 749-773.[21] Slater, L. J. (1966). Generalized Hyper-geometric Functions. *Cambridge University Press*, Cambridge.

Thomson, G. W. (1952). Measures of plant aggregation based on contagious distribution. *Contrib. Lab. Vert. Biol. U. Mich*, 53, pp. 16.

Tripathi, R. C., Gupta, R. C. (1985). A generalization of the log-series distribution, *Communications in Statistics-Theory and Methods*, 14, 1779-1799.

Tripathi, R. C., Gupta, R. C., White, T. J. (1987). Some generalizations of the geometric distribution, *Sankhya, Series B*, 49, 218-223.

doi.org/10.36253/ijas-16754

NON-UNIQUENESS OF $E(s^2)$-OPTIMAL SUPERSATURATED DESIGNS FOR $N \equiv 2$ (MOD 4) RUNS WITH APPLICATION TO THE CASE $N = 10$ RUNS

Francois Domagni[1], **Michael Neubauer**
Department of Mathematics, California State University, Northridge , USA

Abstract *In factor screening experiments with limited resources, it is common for practitioners to cut down the number of runs N and choose a supersaturated design for the experiment. In the past two decades, $E(s^2)$-optimality has been one of the most important criteria used to choose a supersaturated design. We show that the definition $E(s^2)$-optimal supersaturated designs X for $N \equiv 2$ (mod 4) runs and $m \geq N$ factors are not unique by showing that $XX^\top$ allows for multiple non-isomorphic possibilities for most values of m. For $N = 10$ and $12 \leq m \leq 114$ we list all possible $E(s^2)$-optimal designs.*

Keywords: *Supersaturated designs, $E(s^2)$-optimal, Gauss-Dantzig selector, experiments.*

1. INTRODUCTION

Supersaturated designs have gained substantial interest in the last two decades due to their ability to accommodate many more factors than saturated designs in factor screening experiments. Under the sparsity of effect principle, the experimenter may cut down the number of runs required to run a regular fractional factorial design by choosing to run a saturated design. A saturated design may be cost and time saving because it requires as many runs as factors for the experiment (see Domagni et al., 2021). However, in circumstances of limited resources, a supersaturated design is desirable when the number of factors to screen is larger than the number of runs the experimenter can afford. Thus a supersaturated design is a design in which the number of factors m is larger than the number of runs N. A two-level supersaturated design matrix X is a $N \times m$ matrix X with entries from $\{-1,+1\}$ with the assumptions that N is even and $N \leq m$. Furthermore, it is assumed that each column vector x of X is balanced. That is each column vector x has as many $+1$ entries as -1 entries. Lastly, any two columns of X cannot be fully aliased, meaning for two different columns x and y of X, we have $x \neq y$ and $x \neq -y$. It is straightforward from the assumptions that for a given run size N the largest supersaturated design one can construct has $c(N) = \frac{1}{2}\binom{N}{N/2}$ factors.

[1]Francois Domagni kouakou-francois.domagni@csun.edu

In practice it is common that the number of factors m is much smaller than $c(N)$. Define $\text{ssd}(m,N)$ to be the set of all supersaturated designs with run size N and m factors. Under the assumptions underlying supersaturated designs, traditional optimality criteria such D-optimality and A-optimality are not suitable for choosing the best supersaturated designs. Since the column sums of X are 0, $\text{rank}XX^\top < N$, i.e, the $N \times N$ matrix $XX^\top$ is not invertible, and since $\text{rank}(X^\top X) = \text{rank}(XX^\top)$ the $m \times m$ matrix $X^\top X$ is also not invertible. Booth and Cox (1962) proposed the $E(s^2)$-optimal criterion which is by far the most important criterion available in the literature for selecting a supersaturated design for given N and m. See Cheng (1997), Bulutoglu and Cheng (2004), Nguyen (1996), Eskridge et al. (2004). In addition, Jones and Majumdar (2014) propose the $UE(s^2)$-optimal criterion. This criterion does not differ from the $E(s^2)$ criterion except for the fact that the columns of the supersaturated designs need not be balanced. Recently, Singh and Stufken (2023) propose two new design selection criteria inspired by large sample desiderata of the Gauss-Dantzig Selector. These criteria are used with the assumption that only the main effects are active. In this paper, we retain the $E(s^2)$ criterion which measures the average departure from orthogonality of all the pairwise correlation of the m factors given by

$$s^2(X) = \frac{\sum_{i<j} s_{ij}^2}{\binom{m}{2}} \tag{1}$$

where $X^\top X = (s_{ij})$. For a given N and m, the choice of the m column vectors of X that minimizes $s^2(X)$ is the $E(s^2)$-optimal supersaturated design, i.e., we say $X \in \text{ssd}(m,N)$ is $E(s^2)$-optimal if

$$s^2(X) = \min\left\{ s^2(Y) \,|\, Y \in \text{ssd}(m,N) \right\}. \tag{2}$$

It is easy to see that the off-diagonal entries of $X^\top X$ are congruent N modulo 4. Thus $s^2(X) \geq 4$ when $N \equiv 2 \pmod 4$.

It is often easier to calculate the sum of the squares of the off-diagonal elements of $XX^\top = (t_{ij})$ instead. Since

$$\sum_{1\leq i,j\leq m} s_{ij}^2 = \text{trace}X^\top XX^\top X = \text{trace}XX^\top XX^\top = \sum_{1\leq i,j\leq N} t_{ij}^2, \tag{3}$$

$s_{ii} = m$ and $t_{ii} = N$, it follows that

$$\sum_{i\neq j} s_{ij}^2 = \sum_{i\neq j} t_{ij}^2 + mN(m-N). \tag{4}$$

Thus minimizing $\sum_{i\neq j} t_{ij}^2$ is equivalent to minimizing $\sum_{i\neq j} s_{ij}^2$.

In Das et al. (2008) lower bounds for $s^2(X)$ are given that are known to be sharp for all $N \leq m \leq c(N)$ when $N = 6$ and $N = 8$. For $N = 6$ the $E(s^2)$ optimal designs for $6 \leq m \leq 10$ are straightforward to find. E.g., see Domagni and Neubauer (2025). For $N = 8$, Cheng (1997) showed how to construct $E(s^2)$-optimal supersaturated designs for all $8 \leq m \leq 35 = \frac{1}{2}\binom{8}{2}$. They were also given for $m \leq 17$ in Butler et al. (2001). We list those in Domagni and Neubauer (2025) as well. For $N = 10$, $E(s^2)$-optimal supersaturated designs are known for some values of $10 \leq m \leq 126 = \frac{1}{2}\binom{10}{2}$. Cheng (1997) showed how to construct $E(s^2)$-optimal supersaturated designs for $m = 18k \pm 1$ and $18k \pm 2$ where k is a positive integer. In Butler et al. (2001) $E(s^2)$-optimal supersaturated designs are given for $m \leq 22$. However, no full listing of $E(s^2)$-supersaturated designs for all $10 \leq m \leq 126$ was known.

2. A FIRST EXAMPLE OF NON-ISOMORPHIC $E(s^2)$-OPTIMAL SUPERSATURATED DESIGNS

We begin this section by giving some preliminary definitions and properties.

Definition 1. Two supersaturated designs X_1 and $X_2 \in \text{ssd}(m,N)$ are said to be isomorphic if X_2 can be obtained by performing a series of any of these three operations on X_1.

1. Permute any two rows of the matrix X_1.
2. Permute any two columns of X_1.
3. Multiply any column of X_1 by -1.

Recall that a permutation matrix P is a binary square matrix with each line (column or row) contains exactly one 1 and all other entries are 0. Permuting the rows of X_1 is equivalent to pre-multiplying X_1 by the appropriate permutation matrix P. Likewise, permuting the columns of X_1 and multiplying some columns of X_1 by -1 is equivalent to post-multiplying the matrix X_1 by the appropriate matrix Q, where Q is a permutation matrix where some of the non-zero entries may be -1 depending on which columns need to be multiplied by -1. Note that P has dimension N while Q has dimension m.

It follows that two supersaturated designs X_1 and $X_2 \in \text{ssd}(m,N)$ are isomorphic if there exists two matrices P and Q such that $X_2 = PX_1Q$, where P is a permutation matrix and Q is a permutation matrix where the non-zero entries are 1 or -1.

The proof of the next proposition follows directly from the fact that $X_2 = PX_1Q$.

Proposition 2.1. *Suppose X_1 and X_2 are isomorphic, i.e., $X_2 = PX_1Q$, then*

$$X_2X_2^\top = PX_1X_1^\top P^\top \text{ and } X_2^\top X_2 = Q^\top X_1^\top X_1 Q.$$

The entries of the matrix XX^T are the inner products of the rows of X while the entries of $X^\top X$ are the inner products of the columns of X.

Definition 2. Suppose $X \in \text{ssd}(m,N)$ is a supersaturated designs. We say that

1. $XX^\top$ is the row Gram matrix of X and
2. $X^\top X$ is the column Gram matrix of X.

Not that if $X \in \text{ssd}(m,N)$ we can always find a diagonal $\{\pm 1\}$-matrix Q such that all the entries of the first row of XQ are 1. We call such a supersaturated design normalized.

Definition 3. Suppose $X_1, X_2 \in \text{ssd}(m,N)$ are two normalized supersaturated designs. We say that X_1 and X_2 are Gram isomorphic if there exist permutation matrices P and Q such that

$$X_2X_2^\top = PX_1X_1^\top P^\top \text{ and } X_2^\top X_2 = Q^\top X_1^\top X_1 Q.$$

For permutation matrices P and Q we have $P^\top = P^{-1}$ and $Q^\top = Q^{-1}$. Hence we have the following proposition.

Proposition 2.2. *Suppose $X \in ssd(m,N)$ is a normalized supersaturated design. If P is permutation matrix of dimension N and Q is a permutation matrix of dimension m, then $(PXQ)^\top(PXQ) = Q^\top X^\top XQ$ and $(PXQ)(PXQ)^\top = PXX^\top P^\top$.*

It follows immediately that if two normalized supersaturated designs $X_1, X_2 \in \text{ssd}(m,N)$ are isomorphic then they are Gram isomorphic. We will use the converse to show now that there are four non-isomorphic supersaturated designs $X_1, X_2, X_3, X_4 \in \text{ssd}(13,10)$ that pairwise non-isomorphic.

Consider the four normalized supersaturated designs in $\text{ssd}(13,10)$.

$$X_{13,1} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 \\ 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \end{bmatrix}$$

$$X_{13,2} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 \\ -1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 \end{bmatrix}$$

$$X_{13,3} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & -1 \\ -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 \end{bmatrix}$$

$$X_{13,4} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 \\ -1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 \\ -1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 \\ -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \end{bmatrix}$$

They are indeed $E(s^2)$-optimal and, up to isomorphism, are the only $E(s^2)$-optimal designs in $\mathrm{ssd}(13, 10)$, which follows from Proposition 3.5 below. Their row Gram matrices are

$$X_{13,1}X_{13,1}^\top = \left(\begin{array}{ccccc|ccccc} 13 & -3 & -3 & -3 & -3 & 3 & -1 & -1 & -1 & -1 \\ -3 & 13 & 1 & -3 & -3 & -1 & -1 & -1 & -1 & -1 \\ -3 & 1 & 13 & -3 & -3 & -1 & -1 & -1 & -1 & -1 \\ -3 & -3 & -3 & 13 & 1 & -1 & -1 & -1 & -1 & -1 \\ -3 & -3 & -3 & 1 & 13 & -1 & -1 & -1 & -1 & -1 \\ \hline 3 & -1 & -1 & -1 & -1 & 13 & -3 & -3 & -3 & -3 \\ -1 & -1 & -1 & -1 & -1 & -3 & 13 & 1 & -3 & -3 \\ -1 & -1 & -1 & -1 & -1 & -3 & 1 & 13 & -3 & -3 \\ -1 & -1 & -1 & -1 & -1 & -3 & -3 & -3 & 13 & 1 \\ -1 & -1 & -1 & -1 & -1 & -3 & -3 & -3 & 1 & 13 \end{array}\right),$$

$$X_{13,2}X_{13,2}^\top = \left(\begin{array}{ccccc|ccccc} 13 & -3 & -3 & 1 & 1 & -5 & -1 & -1 & -1 & -1 \\ -3 & 13 & 1 & -3 & -3 & -1 & -1 & -1 & -1 & -1 \\ -3 & 1 & 13 & -3 & -3 & -1 & -1 & -1 & -1 & -1 \\ 1 & -3 & -3 & 13 & -3 & -1 & -1 & -1 & -1 & -1 \\ 1 & -3 & -3 & -3 & 13 & -1 & -1 & -1 & -1 & -1 \\ \hline -5 & -1 & -1 & -1 & -1 & 13 & 1 & -3 & 1 & -3 \\ -1 & -1 & -1 & -1 & -1 & 1 & 13 & -3 & -3 & -3 \\ -1 & -1 & -1 & -1 & -1 & -3 & -3 & 13 & -3 & 1 \\ -1 & -1 & -1 & -1 & -1 & 1 & -3 & -3 & 13 & -3 \\ -1 & -1 & -1 & -1 & -1 & -3 & -3 & 1 & -3 & 13 \end{array}\right),$$

$$X_{13,3}X_{13,3}^{\top} = \left(\begin{array}{ccccccc|ccc} 13 & 1 & -3 & -3 & -3 & -3 & 1 & -1 & -1 & -1 \\ 1 & 13 & 1 & -3 & -3 & -3 & -3 & -1 & -1 & -1 \\ -3 & 1 & 13 & 1 & -3 & -3 & -3 & -1 & -1 & -1 \\ -3 & -3 & 1 & 13 & 1 & -3 & -3 & -1 & -1 & -1 \\ -3 & -3 & -3 & 1 & 13 & 1 & -3 & -1 & -1 & -1 \\ -3 & -3 & -3 & -3 & 1 & 13 & 1 & -1 & -1 & -1 \\ 1 & -3 & -3 & -3 & -3 & 1 & 13 & -1 & -1 & -1 \\ \hline -1 & -1 & -1 & -1 & -1 & -1 & -1 & 13 & -3 & -3 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -3 & 13 & -3 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -3 & -3 & 13 \end{array}\right),$$

$$X_{13,4}X_{13,4}^{\top} = \left(\begin{array}{ccc|ccccccc} 13 & -3 & -3 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -3 & 13 & -3 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -3 & -3 & 13 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ \hline -1 & -1 & -1 & 13 & 1 & 1 & -3 & -3 & -3 & -3 \\ -1 & -1 & -1 & 1 & 13 & 1 & -3 & -3 & -3 & -3 \\ -1 & -1 & -1 & 1 & 1 & 13 & -3 & -3 & -3 & -3 \\ -1 & -1 & -1 & -3 & -3 & -3 & 13 & 1 & -3 & 1 \\ -1 & -1 & -1 & -3 & -3 & -3 & 1 & 13 & 1 & -3 \\ -1 & -1 & -1 & -3 & -3 & -3 & -3 & 1 & 13 & 1 \\ -1 & -1 & -1 & -3 & -3 & -3 & 1 & -3 & 1 & 13 \end{array}\right).$$

We first observe that if two normalized supersaturated designs $X_1, X_2 \in \mathrm{ssd}(m,N)$ are isomorphic then the multiset of entries of their row Gram matrices are equal and the multisets of the entries of their column Gram matrices are equal.

Note that 3 is an off-diagonal entry of $X_{13,1}X_{13,1}^{\top}$ but not an off-diagonal entry of the remaining three Gram matrices. This show that $X_{13,1}$ and $X_{13,2}$ are non-isomorphic, $X_{13,1}$ and $X_{13,3}$ are non-isomorphic, and $X_{13,1}$ and $X_{13,4}$ are non-isomorphic.

Likewise, $X_{13,2}$ is not isomorphic to the other three supersaturated designs since $X_{13,2}X_{13,2}^{\top}$ has an off-diagonal entry equal to -5 while none of the other designs have such an entry.

On the other hand, the multisets of entries of $X_{13,3}X_{13,3}^{\top}$ and $X_{13,4}X_{14,4}^{\top}$ are in fact equal. To show that $X_{13,3}$ and $X_{14,4}$ are not isomorphic we consider their

column Gram matrices

$$X_{13,3}^{\top}X_{13,3} = \begin{pmatrix} 10 & 2 & 6 & 2 & -2 & 2 & -2 & 2 & -2 & -2 & 2 & -2 & -2 \\ 2 & 10 & 2 & 2 & 2 & -2 & 2 & -2 & 2 & -2 & 2 & -2 & -2 \\ 6 & 2 & 10 & -2 & 2 & 2 & 2 & 2 & -2 & -2 & -2 & -2 & 2 \\ 2 & 2 & -2 & 10 & 2 & 2 & 2 & 2 & 2 & -2 & 2 & -2 & 2 \\ -2 & 2 & 2 & 2 & 10 & 2 & 2 & 2 & -2 & 2 & -2 & 2 & 2 \\ 2 & -2 & 2 & 2 & 2 & 10 & 2 & -2 & 2 & -2 & -2 & 6 & 2 \\ -2 & 2 & 2 & 2 & 2 & 2 & 10 & -2 & 2 & 2 & -2 & -2 & 2 \\ 2 & -2 & 2 & 2 & 2 & -2 & -2 & 10 & 2 & 2 & -2 & -2 & 2 \\ -2 & 2 & -2 & 2 & -2 & 2 & 2 & 2 & 10 & -2 & -2 & 2 & 2 \\ -2 & -2 & -2 & -2 & 2 & -2 & 2 & 2 & -2 & 10 & 2 & 2 & -2 \\ 2 & 2 & -2 & 2 & -2 & -2 & -2 & -2 & -2 & 2 & 10 & 2 & 2 \\ -2 & -2 & -2 & -2 & 2 & 6 & -2 & -2 & 2 & 2 & 2 & 10 & 2 \\ -2 & -2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & -2 & 2 & 2 & 10 \end{pmatrix}$$

and

$$X_{13,4}^{\top}X_{13,4} = \begin{pmatrix} 10 & 2 & 2 & -2 & -2 & 2 & 2 & -2 & -2 & 2 & -2 & 2 & -2 \\ 2 & 10 & -2 & 2 & 2 & -2 & -2 & -2 & 2 & 2 & 2 & -2 & 2 \\ 2 & -2 & 10 & 2 & 2 & -2 & 2 & -2 & -2 & 2 & 2 & 2 & -2 \\ -2 & 2 & 2 & 10 & 2 & 2 & -2 & 2 & -2 & 2 & -2 & -2 & 2 \\ -2 & 2 & 2 & 2 & 10 & -2 & -2 & 2 & 2 & -6 & 2 & 2 & 2 \\ 2 & -2 & -2 & 2 & -2 & 10 & 2 & 2 & -2 & 2 & 2 & 2 & -2 \\ 2 & -2 & 2 & -2 & -2 & 2 & 10 & -2 & 2 & 2 & 2 & -2 & 2 \\ -2 & -2 & -2 & 2 & 2 & 2 & -2 & 10 & 6 & -2 & -2 & 2 & 2 \\ -2 & 2 & -2 & -2 & 2 & -2 & 2 & 6 & 10 & -2 & 2 & -2 & 2 \\ 2 & 2 & 2 & 2 & -6 & 2 & 2 & -2 & -2 & 10 & 2 & 2 & 2 \\ -2 & 2 & 2 & -2 & 2 & 2 & 2 & -2 & 2 & 2 & 10 & 2 & -2 \\ 2 & -2 & 2 & -2 & 2 & 2 & -2 & 2 & -2 & 2 & 2 & 10 & 2 \\ -2 & 2 & -2 & 2 & 2 & -2 & 2 & 2 & 2 & 2 & -2 & 2 & 10 \end{pmatrix}.$$

We see that $X_{13,4}^{\top}X_{13,4}$ has an entry equal to -6 while while $X_{13,3}^{\top}X_{13,3}$ has no such entry. Hence $X_{13,3}$ and $X_{13,4}$ are not isomorphic.

This finishes the proof that the four $E(s^2)$-optimal supersaturated designs $X_{13,1}, X_{13,2}, X_{13,3}, X_{13,4}$ are not isomorphic.

We comment that since for a permutation matrix P we have $P^{\top} = P^{-1}$ the row (column) Gram matrices of two normalized supersaturated designs are similar and hence have the same eigenvalues. An easy calculation shows that the sets of eigenvalues of the row Gram matrices of $X_{13,1}, X_{13,2}, X_{13,3}, X_{13,4}$ are distinct. This

provides an alternative proof that the four designs are not isomorphic. Given the statistical significance of the off-diagonal entries the proof we gave above is more in the spirit of the subject.

In Section 3 we determine all possible structures of the row Gram matrices of $E(s^2)$-optimal supersaturated designs that achieve the best possible lower bound. In Section 4 we show that for $N = 10$ runs and all m all possible row Gram structures are realized by some $E(s^2)$-optimal supersaturated design in $\mathrm{ssd}(m, 10)$.

3. THE ROW GRAM STRUCTURE OF $E(s^2)$-OPTIMAL SUPERSATURATED DESIGNS

In this section we derive the structure $XX^\top$ for $E(s^2)$-optimal $X \in \mathrm{ssd}(m,N)$. The lower bounds on $E(s^2)$ proved in Das et al. (2008) follow. For each $0 \le b < 2(N-1)$ we determine the structures of $XX^\top$,$X \in \mathrm{ssd}(N, 2(N-1)+b)$, that minimize $\sum_{i\neq j} t_{ij}^2$, which leads to lower bounds for $s^2(X)$ via Equation (4). There are several cases to consider depending on the congruence class of b modulo 4. We write I_k, J_k, 0_k for the identity matrix, the all ones matrix, the zero matrix of size k.

The first result puts constraints on the structure of $XX^\top$. Let $R = (r_{ij}) = XX^\top + 2tJ_N = 2tI_N + P$ when b is even and $R = (r_{ij}) = XX^\top + (2t+1)J_N = 2tI_N + P + J_N$ when b is odd. We state the next result in general but apply it only in the aces $\alpha = 2t$ and $\alpha = 2t+1$.

Lemma 3.1. *Let $X \in ssd(N,m)$. Let $T = (t_{ij}) = XX^\top$ and $R = (r_{ij}) = XX^\top + \alpha J_N$. We have*

1. *$t_{ii} = m$ and $r_{ii} = m + \alpha$ for $1 \le i \le N$,*
2. *$\sum_{i=1}^N t_{ij} = 0$ and $\sum_{j=1}^N r_{ij} = \alpha N$ for $1 \le i \le N$,*
3. *$\sum_{i\neq j} r_{ij} = N((N-1)\alpha - m)$,*
4. *$\sum_{i\neq j} t_{ij}^2 = \sum_{i\neq j} r_{ij}^2 + \alpha N(2m - (N-1)\alpha)$. In particular, $\sum_{i\neq j} t_{ij}^2$ and hence $\sum_{i\neq j} s_{ij}^2$ is minimal if and only if $\sum_{i\neq j} r_{ij}^2$ is minimal.*
5. *There exists a permutation of the rows of X such that*

$$R = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \tag{5}$$

where A is a $p \times p$ matrix, $p \ge \frac{N}{2}$, C a $(N-p)\times(N-p)$ matrix, all elements of A and C are congruent $m+\alpha$ modulo 4 and all entries of B are congruent $m+\alpha+2$ modulo 4.

6. *Lastly, with* $A = (a_{ij})$, $C = (c_{ij})$ *and* $B = (b_{ij})$ *we have*

$$\sum_{i \neq j} c_{ij} - \sum_{k \neq l} a_{kl} = (m - \alpha(N-1))(2p - N). \tag{6}$$

Define

$$\begin{aligned} f(N,t,b) &= (2t(N-1)+b)N(2t(N-1)+b-N) \\ c(N,t,b) &= 4N(N-1) - 4N(2(N-1)+b) \\ g_0(N,t,b) &= 4Nt(t(N-1)+b) + f(N,t,b) \\ g_1(N,t,b) &= N(2t+1)((N-1)(2t-1)+2b) + f(N,t,b). \end{aligned} \tag{7}$$

Below we minimize $\sum_{i \neq j} t_{ij}^2$. The expressions $f(N,t,b)$ accounts for the added term in Equation(4). The expression $c(n,t,b)$ is used in Proposition 3.7. The expressions $g_0(N,t,b)$ respectively $g_1(N,t,b)$ account for the contribution of the term $2tJ_n$ respectively $(2t+1)J_N$ as in part 4 of Lemma 3.1 plus $f(N,t,b)$.

We start with $b \equiv 2 \pmod 4$, the easiest case.

Proposition 3.2. *Let* $b = 4q+2 < N$ *and let* $X \in ssd(N, 2(N-1)t+b)$, $2 \leq b \leq N$, $t \geq 1$. *Then*

$$\sum_{i \neq j} s_{ij}^2 \geq 2N^2 - 4N + 8 + g_0(N,t,b) \tag{8}$$

Furthermore, equality holds if and only if $XX^\top = 2t(NI_N - J_N) + P$ *with*

$$P = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \tag{9}$$

where $A = (b-2)I_{\frac{N}{2}+1} + 2J_{\frac{N}{2}+1} - 4A(G)$, *where* G *is a regular graph of degree* $\frac{N+b}{4}$ *on* $\frac{N}{2}+1$ *vertices,* $C = (b-2)I_{\frac{N}{2}+1} + 2J_{\frac{N}{2}+1} - 4A(H)$, *where* H *is a regular graph of degree* $\frac{N+b-4}{4}$ *on* $\frac{N}{2}-1$ *vertices, while all entries of* B *are 0.*

When $b \equiv 0 \pmod 4$, $0 \leq b < N$, there are 2 cases to consider.

Proposition 3.3. *Let* $b = 4q < N$ *and let* $X \in ssd(N, 2(N-1)t+b)$, $0 \leq b < N$, $t \geq 1$. *Then*

1. *for* $b \leq \frac{N+2}{2}$ *we have*

$$\sum_{i \neq j} s_{ij}^2 \geq 4Nb + g_0(N,t,b) \tag{10}$$

Furthermore, equality holds if and only if

$$P = 4(I_N - A(G)) \tag{11}$$

where G is a regular graph of degree $\frac{b}{4}$ *on N vertices.*

2. *for* $b \geq \frac{N+2}{2}$ *we have*

$$\sum_{i \neq j} s_{ij}^2 \geq 2N^2 + 8b - 8 + g_0(N,t,b) + f(N,t,b) \tag{12}$$

Furthermore, equality holds if and only if $p = \frac{N+2}{2}$ *and*

$$P = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \tag{13}$$

all entries of B have absolute value 2, A has v, off-diagonal entries equal to -4 and C has $\frac{b}{2} - v$ *off-diagonal entries equal to 4 while all other off-diagonal entries of A and C are 0.*

Proposition 3.4. *Let* $b = 4q + 3 < N$ *and let* $X \in ssd(N, 2(N-1)t + b)$, $t \geq 1$. *Then*

1. *for* $b \leq \frac{N}{2}$ *we have*

$$\sum_{i \neq j} s_{ij}^2 \geq 2N^2 + g_1(N,t,b) + f(N,t,b). \tag{14}$$

Furthermore, equality holds if and only if $p = \frac{N}{2}$ *and*

$$P = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} - J_N \tag{15}$$

where all diagonal entries of A and C are equal to $b+1$, *all off-diagonal entries of A and C are 0 and all entries of B have absolute value 2 such that the row and column sums of B are equal to* $N - b - 1$.

2. *for* $b \geq \frac{N}{2}$ *we have*

$$\sum_{i \neq j} s_{ij}^2 \geq 4N^2 - 4bN - 4N + 8b + g_1(N,t,b). \tag{16}$$

Furthermore, equality holds if and only if $p = N-1$ and

$$P = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} - J_N \tag{17}$$

where B has $k = \frac{(3N-2)-(b+1)}{4}$ entries equal to 2 and $N-1-k$ entries equal to -2. Assuming

$$B^\top = (\underbrace{-2,\cdots,-2}_{N-1-k},\underbrace{2,\cdots,2}_{k}) \tag{18}$$

we have

$$A = (m+1)I_{N-1} + 4A(G) \tag{19}$$

where G is a graph with degree sequence $(d^{(N-1-k)},(d-1)^{(k)})$ and $d = k - \frac{N-2}{2}$.

Comment: Matrices B as described in the first part of the proof exist by the Gale-Ryser Theorem (Gale, 1957; Ryser, 1957).

Proposition 3.5. *Let $b = 4q+1 < N$ and let $X \in ssd(N, 2(N-1)t+b)$, $t \geq 1$. Then*

1. *if q is even*

$$\sum_{i \neq j} s_{ij}^2 \geq 2N^2 - 4N + g_{2t+1}(N,t,b). \tag{20}$$

Furthermore, equality holds if and only if $p = \frac{N}{2}$ and

$$P = \begin{pmatrix} (b+1)I_{\frac{N}{2}} + 2J_{\frac{N}{2}} & 0 \\ 0 & (b+1)I_{\frac{N}{2}} + 2J_{\frac{N}{2}} \end{pmatrix} - 4A(G \cup H) - J_{10} \tag{21}$$

where G and H are (disjoint) regular graphs of degree q on $\frac{N}{2}$ vertices.

if q is odd

$$\sum_{i \neq j} s_{ij}^2 \geq 2N^2 - 4N + 32 + g_1(N,t,b) + f(N,t,b) \text{ if } q \text{ is odd} \tag{22}$$

Furthermore, equality holds if and only if

a.

$$P = \begin{pmatrix} (b-1)I_{\frac{N}{2}+2} + 2J_{\frac{N}{2}+2} & 0 \\ 0 & (b-1)bI_{\frac{N}{2}-2} + 2J_{\frac{N}{2}-2} \end{pmatrix} - 4A(G \cup H) - J_N \tag{23}$$

where G is a regular graph of degree $q+1$ on $\frac{N}{2}+2$ vertices and H is a regular graph of degree $q-1$ on $\frac{N-2}{2}$ vertices; or

b.

$$P=\begin{pmatrix}(b+3)I_{\frac{N}{2}}-2J_{\frac{N}{2}} & 0\\ 0 & (b+3)I_{\frac{N}{2}}-2J_{\frac{N}{2}}\end{pmatrix}+4A(G)-J_N \qquad (24)$$

where G is regular graph of degree $d=\frac{N}{2}-1-q$ on $v_1,\cdots,v_N$ such that the induced subgraph G_1 on $v_1,\cdots,v_{\frac{N}{2}}$ has degree sequence $(d^{(\frac{N}{2}-1)},d-1)$ and the induced subgraph G_2 on $v_{\frac{N}{2}+1},\cdots,v_N)$ has degree sequence $(d-1,d^{(\frac{N}{2}-1)})$. The graphs G_1 and G_2 are connected by the edge $(v_{\frac{N}{2}},v_{\frac{N}{2}+1})$; or

c.

$$P=\begin{pmatrix}(b-1)I_{\frac{N}{2}}+2J_{\frac{N}{2}} & 0\\ 0 & b(b-1)I_{\frac{N}{2}}+2J_{\frac{N}{2}}\end{pmatrix}-4A(G)-J_N \qquad (25)$$

where G is regular graph of degree q on $v_1,\cdots,v_N$ such that the induced subgraph G_1 on $v_1,\cdots,v_{\frac{N}{2}}$ has degree sequence $(q^{(\frac{N}{2}-1)},q-1)$ and the induced subgraph G_2 on $v_{\frac{N}{2}+1},\cdots,v_N)$ has degree sequence $(q-1,q^{(\frac{N}{2}-1)})$. The graphs G_1 and G_2 are connected by the edge $(v_{\frac{N}{2}},v_{\frac{N}{2}+1})$

We summarize the results of Propositions 3.2, 3.3, 3.4, and 3.5 first for the case *b* odd and then for the case *b* even.

Proposition 3.6. *Let $X\in ssd(N,2(N-1)t+b)$ with $0\le b<2(N-1)$, b odd, and $N+2\le 2(N-1)+b\le c(N)-(N+2)$. Then*

$$\sum_{i\neq j}s_{ij}^2\ge g_1(N,t,b)+\begin{cases}2N^2 & \text{if} \quad b\equiv 3 \text{ (mod 4) and } b\le\frac{N}{2} \text{ or } b\ge\frac{3N}{2}\\ 8(N-1)+(4N-8)|N-1-b| & \text{if} \quad b\equiv 3 \text{ (mod 4) and } \frac{N}{2}\le b\le\frac{3N}{2}\\ 2N^2-4N & \text{if} \quad b\equiv 1 \text{ (mod 8)}\\ 2N^2-4N+32 & \text{if} \quad b\equiv 5 \text{ (mod 8)}\end{cases}$$

The result for *b* even follows. The proof for *b* even and $b\ge N$ is not complete yet and is given below.

Proposition 3.7. *Let $X \in ssd(N, 2(N-1)t+b)$ with $0 \le b < 2(N-1)$, b even and $N+2 \le 2(N-1)+b \le c(N)-(N+2)$. Let $b' = 2(N-1)-b$. Then*

$$\sum_{i\neq j} s_{ij}^2 \ge g_0(N,t,b) + \begin{cases} 2N^2-4N+8 & \text{if} \quad b \equiv 2 \text{ (mod 4) and } b < N-1 \\ 4Nb'+c(N,t,b) & \text{if} \quad b \equiv 2 \text{ (mod 4) and } b \ge \frac{3N-2}{2} \\ 2N^2+8b'-8+c(N,t,b) & \text{if} \quad b \equiv 2 \text{ (mod 4) and } N \le b \le \frac{3N-2}{2} \\ 4Nb & \text{if} \quad b \equiv 0 \text{ (mod 4) and } b \le \frac{N+2}{2} \\ 2N^2+8b-8 & \text{if} \quad b \equiv 0 \text{ (mod 4) and } \frac{N+2}{2} \le b \le N \\ 2N^2-4N+8+c(N,t,b) & \text{if} \quad b \equiv 0 \text{ (mod 4) and } b \ge N \end{cases}$$

Equality occurs for $b \ge N$ if an only if

$$XX^\top = 2t(NJ_N - J_N) + Q = 2t(NI_N - J_N) + 2(NI_N - J_N) - P$$

where P is as described in

1. *Proposition 3.2 when $b \equiv 0$ (mod 4),*
2. *Part 1 of Proposition 3.3 when $b \equiv 2$ (mod 4) and $b' \le \frac{N+2}{2}$,*
3. *Part 2 of Proposition 3.3 when $b \equiv 2$ (mod 4) and $b' \ge \frac{N+2}{2}$.*

4. THE CASE OF $N = 10$ RUNS

We apply the results of the previous section to the case $N = 10$. We hope this will provide some insight into the complexities of finding all possible structures for $XX^\top$ when X is an $E(s^2)$-optimal supersaturated design for large N.

Corollary 4.1. *If $X \in ssd(10, 18t+2)$, $t \ge 1$, with $\sum_{i\neq j} s_{ij}^2 = 3600t^2 - 1000t + 8$,*

then $XX^\top = 2t(10I_{10} - J_{10}) + P$ *with*

$$P = \begin{pmatrix} 2 & 2 & 2 & -2 & -2 & -2 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & -2 & -2 & -2 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & -2 & -2 & -2 & 0 & 0 & 0 & 0 \\ -2 & -2 & -2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\ -2 & -2 & -2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\ -2 & -2 & -2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & -2 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & -2 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & -2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & -2 & 2 & 2 \end{pmatrix} \text{ or}$$

$$P = \begin{pmatrix} 2 & 2 & -2 & -2 & -2 & 2 & 0 & 0 & 0 & 0 \\ -2 & 2 & 2 & -2 & 2 & 2 & 0 & 0 & 0 & 0 \\ -2 & -2 & 2 & 2 & 2 & -2 & 0 & 0 & 0 & 0 \\ -2 & -2 & -2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 \\ 2 & -2 & -2 & -2 & 2 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & -2 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & -2 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & -2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & -2 & 2 & 2 \end{pmatrix} \tag{26}$$

There exist examples that realize these two Gram structures for all $1 \le t \le 7$ *which we give in Section 6.*

If $X \in ssd(10, 18t+6)$, $t \ge 1$, *with* $\sum_{i \ne j} s_{ij}^2 = 3600t^2 + 600t - 72$, *then* $XX^\top = 2t(10I_{10} - J_{10}) + P$ *with*

$$P = \begin{pmatrix} 6 & 2 & -2 & -2 & -2 & -2 & 0 & 0 & 0 & 0 \\ 2 & 6 & -2 & -2 & -2 & -2 & 0 & 0 & 0 & 0 \\ -2 & -2 & 6 & 2 & -2 & -2 & 0 & 0 & 0 & 0 \\ -2 & -2 & 2 & 6 & -2 & -2 & 0 & 0 & 0 & 0 \\ -2 & -2 & -2 & -2 & 6 & 2 & 0 & 0 & 0 & 0 \\ -2 & -2 & -2 & -2 & 2 & 6 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 6 & -2 & -2 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & 6 & -2 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & -2 & 6 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & -2 & -2 & 6 \end{pmatrix} \tag{27}$$

In particular, there is up to Gram isomorphism only one $E(s^2)$*-optimal supersaturated design for all* $1 \le t \le 7$ *which we give in Section 6.*

If $X \in ssd(10, 18t+10)$, $t \geq 1$, with $\sum_{i \neq j} s_{ij}^2 = 3600t^2 + 2280t + 296$, then $XX^\top = (2t+1)(10I_{10} - J_{10}) - P$ with with P as in Equation (29). There exist examples that realize these four Gram structures for all $1 \leq t \leq 6$ which we give in Section 6.

If $X \in ssd(10, 18t+14)$, $t \geq 0$, with $\sum_{i \neq j} s_{ij}^2 = 3600t^2 + 4200t + 920$, then $XX^\top = (2t+1)(10I_{10} - J_{10}) - P$ with P as in Equation (28). In particular there is up to Gram isomorphism only one $E(s^2)$-optimal supersaturated design for each $1 \leq t \leq 6$ which we give in Section 6.

Since $s(X^2) \geq 4$ the bound of Corollary 4.1 does not apply when $m = 10$. In fact, $s(X^2) = 4$ for $E(s^2)$-optimal supersaturated designs $X \in \mathrm{ssd}(10, 10)$.

Corollary 4.2. *If $X \in ssd(10, 18t)$, $t \geq 1$, with $\sum_{i \neq j} s_{ij}^2 = 3600t^2 - 1800t$, then $XX^\top = 2t(10I_{10} - J_{10})$, i.e., $P = 0_N$.*

If $X \in ssd(10, 18t+4)$, $t \geq 1$, with $\sum_{i \neq j} s_{ij}^2 = 3600t^2 - 200t - 80$, then $XX^\top = 2t(10I_{10} - J_{10}) + P$ with

$$P = \begin{pmatrix} 4 & -4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -4 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & -4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -4 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & -4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -4 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & -4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -4 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 & -4 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -4 & 4 \end{pmatrix}. \tag{28}$$

In particular there is up to Gram isomorphism only one $E(s^2)$-optimal supersaturated design for each $1 \leq t \leq 7$ which we give in Section 6.

If $X \in ssd(10, 18t+8)$, $t \geq 1$, with $\sum_{i \neq j} s_{ij}^2 = 3600t^2 + 1400t + 96$, then $XX^\top = 2t(10I_{10} - J_{10}) + P$ with

$$
P=\left(\begin{array}{cccccc|cccc}
8 & 0 & 0 & 0 & 0 & 0 & -2 & -2 & -2 & -2\\
0 & 8 & 0 & 0 & 0 & 0 & -2 & -2 & -2 & -2\\
0 & 0 & 8 & -4 & 0 & 0 & -2 & -2 & 2 & -2\\
0 & 0 & -4 & 8 & 0 & 0 & -2 & 2 & -2 & -2\\
0 & 0 & 0 & 0 & 8 & -4 & 2 & -2 & -2 & -2\\
0 & 0 & 0 & 0 & -4 & 8 & -2 & -2 & -2 & 2\\
\hline
-2 & -2 & -2 & -2 & 2 & -2 & 8 & 0 & 0 & 0\\
-2 & -2 & -2 & 2 & -2 & -2 & 0 & 8 & 0 & 0\\
-2 & -2 & 2 & -2 & -2 & -2 & 0 & 0 & 8 & 0\\
-2 & -2 & -2 & -2 & -2 & 2 & 0 & 0 & 0 & 8
\end{array}\right) \quad or
$$

$$
P=\left(\begin{array}{cccccc|cccc}
8 & 0 & 0 & 0 & 0 & 0 & -2 & -2 & -2 & -2\\
0 & 8 & 0 & 0 & 0 & 0 & -2 & -2 & -2 & -2\\
0 & 0 & 8 & 0 & 0 & 0 & -2 & -2 & -2 & -2\\
0 & 0 & 0 & 8 & 0 & 0 & -2 & -2 & -2 & -2\\
0 & 0 & 0 & 0 & 8 & 0 & -2 & -2 & -2 & -2\\
0 & 0 & 0 & 0 & 0 & 8 & -2 & -2 & -2 & -2\\
\hline
-2 & -2 & -2 & -2 & -2 & -2 & 8 & 0 & 4 & 0\\
-2 & -2 & -2 & -2 & -2 & -2 & 0 & 8 & 0 & 4\\
-2 & -2 & -2 & -2 & -2 & -2 & 4 & 0 & 8 & 0\\
-2 & -2 & -2 & -2 & -2 & -2 & 0 & 4 & 0 & 8
\end{array}\right) \quad or
$$

$$
P=\left(\begin{array}{cccccc|cccc}
8 & -4 & 0 & 0 & 0 & 0 & -2 & 2 & -2 & -2\\
-4 & 8 & 0 & 0 & 0 & 0 & 2 & -2 & -2 & -2\\
0 & 0 & 8 & 0 & 0 & 0 & -2 & -2 & -2 & -2\\
0 & 0 & 0 & 8 & 0 & 0 & -2 & -2 & -2 & -2\\
0 & 0 & 0 & 0 & 8 & 0 & -2 & -2 & -2 & -2\\
0 & 0 & 0 & 0 & 0 & 8 & -2 & -2 & -2 & -2\\
\hline
-2 & 2 & -2 & -2 & -2 & -2 & 8 & 0 & 0 & 0\\
2 & -2 & -2 & -2 & -2 & -2 & 0 & 8 & 0 & 0\\
-2 & -2 & -2 & -2 & -2 & -2 & 0 & 0 & 8 & 4\\
-2 & -2 & -2 & -2 & -2 & -2 & 0 & 0 & 4 & 8
\end{array}\right) \quad or
$$

$$
P=\left(\begin{array}{cccccc|cccc}
8 & -4 & -4 & 0 & 0 & 0 & 2 & 2 & -2 & -2\\
-4 & 8 & 0 & 0 & 0 & 0 & -2 & -2 & -2 & 2\\
-4 & 0 & 8 & 0 & 0 & 0 & -2 & -2 & 2 & -2\\
0 & 0 & 0 & 8 & 0 & 0 & -2 & -2 & -2 & -2\\
0 & 0 & 0 & 0 & 8 & 0 & -2 & -2 & -2 & -2\\
0 & 0 & 0 & 0 & 0 & 8 & -2 & -2 & -2 & -2\\
\hline
2 & -2 & -2 & -2 & -2 & -2 & 8 & 0 & 0 & 0\\
2 & -2 & -2 & -2 & -2 & -2 & 0 & 8 & 0 & 0\\
-2 & -2 & 2 & -2 & -2 & -2 & 0 & 0 & 8 & 0\\
-2 & 2 & -2 & -2 & -2 & -2 & 0 & 0 & 0 & 8
\end{array}\right).
$$

(29)

There exist examples that realize these four Gram structures for all $1 \leq t \leq 7$ *which we give in Section 6.*

If $X \in ssd(10, 18t+12)$, $t \geq 0$, *with* $\sum_{i \neq j} s_{ij}^2 = 3600t^2 + 3240t + 528$, *then* $XX^\top = (2t+1)(10I_{10} - J_{10}) - P$ *with P as in Equation (27). In particular there is up to Gram isomorphism only one* $E(s^2)$*-optimal supersaturated design for each* $1 \leq t \leq 6$ *which we give in Section 6.*

If $X \in ssd(10, 18t+16)$, $t \geq 0$, *with* $\sum_{i \neq j} s_{ij}^2 = 3600t^2 + 5160t + 1408$, *then* $XX^\top = (2t+1)(10I_{10} - J_{10}) - P$ *with P as in Equation (26). There exist examples that realize these two Gram structures for all* $1 \leq t \leq 6$ *which we give in Section 6.*

Corollary 4.3. *If* $X \in ssd(10, 18t+3)$, $t \geq 1$, *with* $\sum_{i \neq j} s_{ij}^2 = 3600t^2 - 600t - 40$, *then* $XX^\top = 2t(10I_{10} - J_{10}) + P$ *with*

$$P = \left(\begin{array}{rrrrr|rrrrr} 3 & -1 & -1 & -1 & -1 & -3 & 1 & 1 & 1 & 1 \\ -1 & 3 & -1 & -1 & -1 & 1 & -3 & 1 & 1 & 1 \\ -1 & -1 & 3 & -1 & -1 & 1 & 1 & -3 & 1 & 1 \\ -1 & -1 & -1 & 3 & -1 & 1 & 1 & 1 & -3 & 1 \\ -1 & -1 & -1 & -1 & 3 & 1 & 1 & 1 & 1 & -3 \\ \hline 3 & 1 & 1 & 1 & 1 & 3 & -1 & -1 & -1 & -1 \\ 1 & -3 & 1 & 1 & 1 & -1 & 3 & -1 & -1 & -1 \\ 1 & 1 & -3 & 1 & 1 & -1 & -1 & 3 & -1 & -1 \\ 1 & 1 & 1 & -3 & 1 & -1 & -1 & -1 & 3 & 1 \\ 1 & 1 & 1 & 1 & -3 & -1 & -1 & -1 & -1 & 3 \end{array}\right). \tag{30}$$

There exist examples that realize this Gram structure. In particular there is up to Gram isomorphism only one $E(s^2)$*-optimal supersaturated design for each* $1 \leq t \leq 6$ *which we give in Section 6.*

If $X \in ssd(10, 18t+7)$, $t \geq 1$, *with* $\sum_{i \neq j} s_{ij}^2 = 3600t^2 + 1000t - 24$, *then* $XX^\top = 2t(10I_{10} - J_{10}) + P$ *with*

$$P = \left(\begin{array}{cccc|cccccc} 8 & 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 \\ 4 & 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -2 \\ 0 & 0 & 8 & 4 & 0 & 0 & 0 & 0 & 0 & -2 \\ 0 & 0 & 4 & 8 & 0 & 0 & 0 & 0 & 0 & -2 \\ \hline 0 & 0 & 0 & 0 & 8 & 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 8 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 8 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 8 & 2 \\ \hline -2 & -2 & -2 & -2 & 2 & 2 & 2 & 2 & 2 & 8 \end{array}\right) - J_{10}. \tag{31}$$

There exist examples that realize this Gram structure. In particular, there is up to Gram isomorphism only one $E(s^2)$-optimal supersaturated design for each $1 \le t \le 6$ which we give in Section 6.

If $X \in ssd(10, 18t+11)$, $t \ge 1$, with $\sum_{i \ne j} s_{ij}^2 = 3600t^2 + 2600t + 376$, then $XX^\top = (2t+1)(10I_{10} - J_{10}) - P$ with P as in Equation (31). In particular, there is up to Gram isomorphism only one $E(s^2)$-optimal supersaturated design for each $1 \le t \le 6$ which we give in Section 6.

If $X \in ssd(10, 18t+15)$, $t \ge 0$, with $\sum_{i \ne j} s_{ij}^2 = 3600t^2 + 4680t + 1400$, then $XX^\top = (2t+1)(10I_{10} - J_{10}) - P$ with P as in Equation (30). In particular there is up to Gram isomorphism only one $E(s^2)$-optimal supersaturated design for each $1 \le t \le 6$ which we give in Section 6.

Since $s(X^2) \ge 4$ the bound of Corollary 4.3 does not apply when $m = 11$. In fact, $s(X^2) = 4$ for $E(s^2)$-optimal supersaturated designs $X \in \mathrm{ssd}(11, 10)$.

Corollary 4.4. *If $X \in ssd(10, 18t+1)$, $t \ge 1$, with $\sum_{i \ne j} s_{ij}^2 = 3600t^2 - 1400t$, then $XX^\top = 2t(10I_{10} - J_{10}) + P$ with*

$$P = \left(\begin{array}{ccccc|ccccc} 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 \\ \hline -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \end{array}\right). \tag{32}$$

There exist examples that realize this Gram structure. See Section 6. In particular there is up to Gram isomorphism only one $E(s^2)$-optimal supersaturated design for each $1 \leq t \leq 7$ which we give in Section 6.

If $X \in ssd(10, 18t+5)$, $t \geq 1$, with $\sum_{i \neq j} \sum_{i \neq j} s_{ij}^2 = 3600t^2 + 200t - 48$, then $XX^\top = 2t(10I_{10} - J_{10}) + P$ with

$$P = \left(\begin{array}{rrrrrrr|rrr} 6 & -2 & 2 & 2 & 2 & 2 & -2 & 0 & 0 & 0 \\ -2 & 6 & -2 & 2 & 2 & 2 & 2 & 0 & 0 & 0 \\ 2 & -2 & 6 & -2 & 2 & 2 & 2 & 0 & 0 & 0 \\ 2 & 2 & -2 & 6 & -2 & 2 & 2 & 0 & 0 & 0 \\ 2 & 2 & 2 & -2 & 6 & -2 & 2 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & -2 & 6 & -2 & 0 & 0 & 0 \\ -2 & 2 & 2 & 2 & 2 & -2 & 6 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 6 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 6 \end{array}\right) - J_{10} \text{ or}$$

$$P = \left(\begin{array}{rrrrrrr|rrr} 6 & -2 & -2 & 2 & 2 & 2 & 2 & 0 & 0 & 0 \\ -2 & 6 & -2 & 2 & 2 & 2 & 2 & 0 & 0 & 0 \\ -2 & -2 & 6 & 2 & 2 & 2 & 2 & 0 & 0 & 0 \\ 2 & 2 & 2 & 6 & -2 & 2 & -2 & 0 & 0 & 0 \\ 2 & 2 & 2 & -2 & 6 & -2 & 2 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & -2 & 6 & -2 & 0 & 0 & 0 \\ 2 & 2 & 2 & -2 & 2 & -2 & 6 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 6 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 6 \end{array}\right) - J_{10} \text{ or}$$

$$P = \left(\begin{array}{rrrrr|rrrrr} 6 & -2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ -2 & 6 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 6 & 2 & -2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 6 & -2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & -2 & -2 & 6 & 4 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 4 & 6 & -2 & -2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & -2 & 6 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & -2 & 2 & 6 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 6 & -2 \\ 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & -2 & 6 \end{array}\right) - J_{10} \text{ or}$$

$$P = \left(\begin{array}{rrrrr|rrrrr} 6 & -2 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ -2 & 6 & 2 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 6 & -2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & -2 & 6 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 2 & 2 & 2 & 6 & -4 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & -4 & 6 & 2 & 2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 2 & 6 & -2 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 2 & -2 & 6 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & 6 & -2 \\ 0 & 0 & 0 & 0 & 0 & 2 & 2 & 2 & -2 & 6 \end{array}\right) - J_{10}. \quad (33)$$

There exist examples that realize these four Gram structures for all $1 \leq t \leq 7$ *which we give in Section 6.*

If $X \in ssd(10, 18t+9)$, $t \geq 1$, *with* $\sum_{i \neq j} s_{ij}^2 = 3600t^2 + 1800t + 160$, *then* $XX^\top = 2t(10I_{10} - J_{10}) + P$ *with*

$$P = \left(\begin{array}{rrrrrrrrrr} 10 & 2 & -2 & -2 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 10 & 2 & -2 & -2 & 0 & 0 & 0 & 0 & 0 \\ -2 & 2 & 10 & 2 & -2 & 0 & 0 & 0 & 0 & 0 \\ -2 & -2 & 2 & 10 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & -2 & -2 & 2 & 10 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 10 & 2 & -2 & -2 & 2 \\ 0 & 0 & 0 & 0 & 0 & 2 & 10 & 2 & -2 & -2 \\ 0 & 0 & 0 & 0 & 0 & -2 & 2 & 10 & 2 & -2 \\ 0 & 0 & 0 & 0 & 0 & -2 & -2 & 2 & 10 & 2 \\ 0 & 0 & 0 & 0 & 0 & 2 & -2 & -2 & 2 & 10 \end{array}\right) - J_{10} \quad (34)$$

There exist examples that realize this Gram structure. See Section 6. In particular there is up to Gram isomorphism only one $E(s^2)$*-optimal supersaturated design for each* $1 \leq t \leq 6$ *which we give in Section 6.*

If $X \in ssd(10, 18t+13)$, $t \geq 0$, *with* $\sum_{i \neq j} \sum_{i \neq j} s_{ij}^2 = 3600t^2 + 3720t + 912$, *then* $XX^\top = 2t(10I_{10} - J_{10}) + P$ *with P as in Equation (33). There exist examples that realize these Gram structures for all* $1 \leq t \leq 6$ *which we give in Section 6.*

If $X \in ssd(10, 18t+17)$, $t \geq 0$, *with* $\sum_{i \neq j} \sum_{i \neq j} s_{ij}^2 = 3600t^2 + 5640t + 1920$, *then* $XX^\top = 2t(10I_{10} - J_{10}) + P$ *with P as in Equation (32). In particular there is up to Gram isomorphism only one* $E(s^2)$*-optimal supersaturated design for each* $1 \leq t \leq 6$ *which we give in Section 6.*

5. THE PROOFS

Let $X \in \mathrm{ssd}(N, 2t(N-1)+b)$, $0 \le b < N$. We write $T = (t_{ij}) = XX^\top = 2t(NI_n - J_N) + P$, $P = (p_{ij})$. Since $t_{ii} = 2t(N-1)+b$ and $\sum_{j=1}^{N} t_{ij} = 0$ for $1 \le i \le N$, it follows that $p_{ii} = b$ and $\sum_{j=1}^{N} p_{ij} = 0$ for $1 \le i \le N$.

Proof of Lemma 3.1. The results follow by straightforward calculations from the facts that $t_{ii} = m$ and $\sum_{j=1}^{N} t_{ij} = 0$ for $1 \le i \le N$. □

Proof of Proposition 3.2. We use $\alpha = 2t$ and assume $R = XX^\top + 2tJ_N$ is partitioned as in Lemma 3.1 with $p \ge \frac{N}{2}$ even. Notice that $r_{ii} \equiv 2 \pmod 4$. Thus all off-diagonal entries of A and C have absolute value at least 2. Thus

$$\sum_{i \ne j} r_{ij}^2 \ge 4p(p-1) + 4(N-p)(N-p-1) = 8p(p-n) + 4n(n-1) = w(p). \quad (35)$$

Since $p \ge \frac{N}{2}$ is even $w(p)$ takes on its minimal value at $p = \frac{N}{2} + 1$. Since the row sum of A and C are 0, we find each row of A has exactly $\frac{N+b}{4}$ entries equal to -2 and each row of C has exactly $\frac{N+b-4}{4}$ entries equal to -2. This allows us to write and A and C as stated. □

Proof of Corollary 4.1.

1. If $b = 2$, Proposition 3.2 states that G is a regular graph of degree 3 on 6 vertices and H is a regular graph of degree 2 on 4 vertices. This implies that $G = K_3 \oplus K_3$ or $G = K_6 - C_6$ while $H = C_4$. This yields the two matrices for P.

2. If $b = 6$, Proposition 3.2 states that G is a regular graph of degree 4 on 6 vertices and H is a regular graph of degree 3 on 4 vertices. Hence $G = (K_2 \cup K_2 \cup K_2)^2$ and $H = K_3$. This determines P.

□

Proof of Proposition 3.3. We use $\alpha = 2t$ and assume $R = (r_{ij}) = XX^\top + 2tJ_N$ is partitioned as in Lemma 3.1 with $p \ge \frac{N}{2}$ even. Notice that all entries of B have absolute value at least 2. Furthermore, since $\sum_{i \ne j} c_{ij} - \sum_{k \ne l} a_{kl} = (2p-N)b$ and all off-diagonal entries of A and C are congruent 0 modulo 4, we have

$$\sum_{i \ne j} c_{ij}^2 + \sum_{k \ne l} a_{kl}^2 \ge 4(2p-N)b. \quad (36)$$

Thus

$$\sum_{i \ne j} r_{ij}^2 \ge 8p(N-p) + 4(2p-N)b = -8p^2 + 8(N+b)p - 4Nb = w(p). \quad (37)$$

The function $w(p)$ has a vertex at $p = \frac{N}{2} + \frac{b}{2}$. The range of p is $[\frac{N}{2}+1, N]$, which has its midpoint at $\frac{3N+2}{4}$.

If $b \leq \frac{N}{2}+1$ the minimum of $d(p)$ occurs at the right endpoint N and we can write $XX^\top = 2t(NI_N - J_N) - 4A(G)$ where G is a regular graph of degree $\frac{b}{4}$ on N vertices.

If $b \geq \frac{N}{2}+1$ the minimum of $d(p)$ occurs at the left endpoint N. There are several possibilities for the minimum to occur. All entries of B have absolute value 2, A has v, off-diagonal entries equal to -4 and C has $\frac{b}{2} - v$ off-diagonal entries equal to 4 while all other off-diagonal entries of A and C are 0. □

Proof of Corollary 4.2. 1. If $b = 0$, Proposition 3.3 states that $P = 0_N$.

2. If $b = 4$, Proposition 3.3 states that G is a regular graph of degree 1 on 10 vertices. Hence $G = K_2 \cup K_2 \cup K_2 \cup K_2 \cup K_2$. This determines P.

3. If $b = 8$, Proposition 3.3 states that $p = 6$ and hence C has either 0,2 or 4 entries equal to 4. Since $p = 6$ and each entry of B has absolute value 2, each row and column of C has at most one off-diagonal entry equal to 4. This leads to the four possible structures for P as listed.

□

For the case $b \equiv 3 \pmod 4$ we use Lemma 3.1 with $\alpha = 2t+1$.

Proof of Proposition 3.4. We use $\alpha = 2t+1$ and assume $R = (r_{ij}) = XX^\top + (2t+1)J_N$ is partitioned as in Lemma 3.1 with $p \geq \frac{N}{2}$ odd. Notice that all entries of B have absolute value at least 2. Furthermore, since $\sum_{i\neq j} c_{ij} - \sum_{k\neq l} a_{kl} = (2p-N)(b-N+1)$ and all off-diagonal entries of A and C are congruent 0 modulo 4, we have

$$\sum_{i\neq j} c_{ij}^2 + \sum_{k\neq l} a_{kl}^2 \geq 4(2p-N)|N-1-b| \tag{38}$$

and hence

$$\sum_{i\neq j} r_{ij}^2 \geq 8p(N-p) + 4(2p-N)|N-1-b| = w(p). \tag{39}$$

The range of p is $[\frac{N}{2}, N-1]$, which has its midpoint at $\frac{3N-2}{4}$.

If $b \leq N-1$ then $w(p)$ has a vertex at $N - \frac{b+1}{2}$. Hence, if $N - \frac{b+1}{2} \geq \frac{3N+2}{4}$, i.e., $b \leq \frac{N}{2}$, then the minimum of $f(p)$ occurs at the left endpoint $\frac{N}{2}$. In that case, all off-diagonal entries of A and C are 0 and each row and column of B has $\frac{b+1}{4}$

entries equal to -2 and $\frac{N}{2} - \frac{b+1}{4}$ entries equal to 2. It follows that $\sum_{i\neq j} r_{i,j} = 2N^2$ and hence $\sum_{i\neq j} s_{ij} = 2N^2 + g_1(N,t,b)$.

Likewise, if $b \geq \frac{N}{2}$, then the minimum of $f(p)$ occurs at the right endpoint $N-1$. In that case B has $k = \frac{(3N-2)-(b+1)}{4}$ entries equal to 2 and $N-1-k$ entries equal to -2. Assuming

$$B^\top = (\underbrace{-2,\cdots,-2}_{N-1-k},\underbrace{2,\cdots,2}_{k}) \tag{40}$$

we have

$$A = (m+1)I_{N-1} + 4A(G) \tag{41}$$

where G is a graph with degree sequence $(d^{(N-1-k)},(d-1)^{(k)})$ and $d = k - \frac{N-2}{2}$. The expression for $\sum_{i\neq j} s_{ij}^2$ follows.

If $b > N-1$, then $w(p)$ has a vertex at $\frac{b+1}{2}$. As above, if $b \geq \frac{3N-2}{4}$, then the minimum occurs at the left end point of the interval $[\frac{N}{2}, N-1]$ and the result follows as above.

If $N-1 < b \leq \frac{3N-2}{4}$, then the minimum of $w(p)$ occurs at the right endpoint $N-1$ and we proceed as above. □

Proof of Corollary 4.3. 1. If $b = 3$, Proposition 3.4 implies that there is a unique solution

$$P = (4I_5 - J_5) \otimes \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \tag{42}$$

2. If $b = 7$, Proposition 3.4 states that $p = 9$ and

$$P = \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} - J_{10} \tag{43}$$

such that B has five entries equal to 2 and four entries equal to -2 and $A = 8I_9 + 4A(G)$ where G is a graph with degree sequence $(1,1,1,1,0,0,0,0,0)$. Thus P is a as described.

□

For the case $b \equiv 1 \pmod 4$ we use Lemma 3.1 with $\alpha = 2t+1$. It splits into 2 cases, $b = 8q+1$ and $b = 8q+5$, the latter of which leads to 3 possibilities for the Gram structure of $XX^\top$.

Proof of Proposition 3.5. We use $\alpha = 2t+1$ and assume $R = (r_{ij}) = XX^\top + (2t+1)J_N$ is partitioned as in Lemma 3.1 with $p \geq \frac{N}{2}$ odd. Since all off-diagonal entries of A and C have absolute value at least 2 we have

$$\sum_{i \neq j} r_{ij}^2 \geq 4p(p-1) + 4(N-p)(N-p-1) = w(p). \tag{44}$$

The function $w(p)$ has a minimum at $\frac{N}{2}$.

If $b = 8q+1$ then the minimum is attained by the example as stated above.

If $b = 8q+5$ is odd and $p = \frac{N}{2}$, then at least one entry of B has absolute value at least 4. In that case,

$$\sum_{i \neq j} r_{ij}^2 \geq 2N(N-2) + 32. \tag{45}$$

This bound is achievable by the examples stated above. This bound can also be achieved for $p = \frac{N}{2} + 2$ and all entries of B equal to 0, which leads to examples as stated above. □

Proof of Corollary 4.4.

1. If $b = 1$, Proposition 3.5 states that G and H are regular graphs of degree 0 on 5 vertices. Thus G and H are graphs with no vertices, i.e. $A(G \cup H)$ is the zero matrix. This determines P.

2. If $b = 9$, Proposition 3.5 states that G and H are regular graphs of degree 2 on 5 vertices. Thus $G = C_5 = H$. This determines P.

3. If $b = 5$, Proposition 3.5 states that

   (a) if $p = \frac{N}{2} + 2$, then G is a regular graph of degree 2 on 7 vertices and H is a regular graph of degree 0 on 3 vertices. Thus $G = C_7$ or $G = C_3 \cup C_4$. This gives rise to the first two examples. OR

   (b) G is a regular graph of degree 3 on 10 vertices such that the induced subgraph on vertices v_1, v_2, v_3, v_4, v_5 has degree sequence $(3,3,3,3,2)$ and the induced subgraph on vertices $v_6, v_7, v_8, v_9, v_{10}$ has degree sequence $(2,3,3,3,3)$. OR

   (c) G is a regular graph of degree 1 on 10 vertices such that the induced subgraph on vertices v_1, v_2, v_3, v_4, v_5 has degree sequence $(1,1,1,1,0)$ and the induced subgraph on vertices $v_6, v_7, v_8, v_9, v_{10}$ has degree sequence $(0,1,1,1,1)$.

   Thus P is one of the matrices listed.

□

Proof of Proposition 3.7. Write $XX^\top = 2t(NI_N - J_N) + Q = 2t(NI_N - J_N) + (2NI_N - 2J_N - P)$. Notice that

$$\begin{aligned} p_{ii} &= 2(N-1) - b, \\ \sum_{j \neq i} p_{ij} &= -2(N-1) + b, \\ \sum_{i \neq j} q_{ij}^2 &= 4N(b - (N-1)) + \sum_{i \neq j} p_{ij}^2. \end{aligned} \tag{46}$$

Thus minimizing $\sum_{i \neq j} q_{ij}^2$ is equivalent to minimizing $\sum_{i \neq j} p_{ij}^2$. Furthermore, P satisfies the conditions of Propositions 3.2, 3.3 for the respective $0 \leq b' = 2(N-1) - b < N$. □

6. THE EXAMPLES FOR $N = 10$ RUNS

Let $N = 2k$ and $C(2k,k)$ be the matrix whose columns are the (± 1)-vectors of length $2k$ with k entries equal to -1 and first entry equal to 1 ordered lexicographically. For example, the first column of $C(10,5)$ is $(1,1,1,1,1,-1,-1,-1,-1,-1)^\top$, the second column is $(1,1,1,1,-1,1,-1,-1,-1,-1)^\top$ and the 126th column is $(1,-1,-1,-1,-1,-1,1,1,1,1)^\top$. As such each column has an address and we list designs by the addresses of their columns.

We list supersaturated designs by giving the list of column addresses. For example, we write $X_{12} = (17,20,25,31,40,57,71,76,98,105,121,124)$ for the supersaturated design with $N = 10$ runs and $m = 12$ factors whose columns are column 17 of $C(10,5)$, column 20 of $C(10,5)$ and so on.

We list the examples for values $12 \leq m \leq 63$. For $m = 10, 11$ the bounds of Propositions 3.6 and 3.7 do not apply. For $63 < m \leq 114$ we obtain $E(s^2)$-optimal supersaturated designs via complementation of $E(s^2)$-optimal supersaturated designs for $12 \leq 126 - m < 63$ factors as the next result demonstrates.

Lemma 6.1. *Let $c = |C(N,N/2)|$. Assume $X_1 \in ssd(m,N)$, $N \leq m \leq c - N$ and X_1 is $E(s^2)$-optimal. If $X_2 \in ssd(c-m,N)$ is the design formed by all columns in $C(N,N/2)$ which are not in X_1, then X_2 is $E(s^2)$-optimal.*

Proof. Let $X \in \mathrm{ssd}(m,N)$ and let $Y \in \mathrm{ssd}(b-m,N)$ be the complement of X in $B(N)$. Note that

$$XX^\top = [X,Y]\,[X,Y]^\top = XX^\top + YY^\top = sI_N - \frac{s}{N} J_N. \tag{47}$$

with $s = \frac{Nc}{(N-1)}$. Hence

$$YY^\top = sI_N - \frac{s}{N}J_N - XX^\top. \tag{48}$$

With $XX^\top = (t_{ij})$ and $YY^\top = (\tau_{ij})$ we have

$$\tau_{ij} = -\frac{c}{N-1} - t_{ij} \text{ for } i \neq j \tag{49}$$

This implies

$$\sum_{i\neq j}(\tau_{ij})^2 = \sum_{i\neq j}\left(-\frac{c}{N-1} - t_{ij}\right)^2 = \frac{cN(c-2m)}{N-1} + \sum_{i\neq j} t_{ij}^2. \tag{50}$$

Thus X is $E(s^2)$-optimal if and only if Y is $E(s^2)$-optimal. □

6.1. The examples for $b \equiv 2$ (mod 4)

$X_{14} = (2,21,25,31,44,49,57,70,76,83,98,105,110,116)$

$X_{20,1} = (7,14,16,18,21,25,27,31,45,56,59,68,69,77,86,95,107,112,120,122)$

$X_{20,2} = (3,11,14,21,24,27,38,50,54,55,65,69,76,82,95,97,111,113,117,123)$

$X_{24} = (4,16,21,25,26,31,38,39,48,49,62,64,73,75,77,78,91,93,98,103,111,118,123,125)$

$X_{28,1} = (2,18,19,21,23,31,34,39,43,45,47,52,58,61,63,67,73,79,83,89,95,106,110,111,113,117,118,123)$

$X_{28,2} = (15,16,18,20,23,28,34,36,37,44,46,48,60,64,65,67,73,76,81,88,98,100,102,109,114,115,124,125)$

$X_{28,3} = (3,6,14,19,20,25,32,35,38,39,42,43,49,51,61,64,75,76,81,82,88,93,96,104,105,111,121,124)$

$$X_{28,4} = (6, 12, 15, 17, 20, 29, 32, 36, 37, 38, 41, 46, 57, 59, 69, 75, 77, 78, 85, 88, 94, 98, 102, 104, 109, 121, 124, 125)$$

$$X_{32} = X_{14} \cup (1, 15, 19, 32, 33, 39, 45, 47, 58, 63, 81, 84, 88, 93, 107, 111, 113, 119)$$

$$X_{38,1} = X_{20,1} \cup (1, 12, 20, 32, 36, 40, 41, 50, 60, 63, 78, 83, 91, 93, 104, 109, 117, 118)$$

$$X_{38,2} = X_{20,2} \cup (1, 15, 17, 34, 35, 39, 43, 47, 59, 62, 81, 86, 88, 92, 107, 110, 115, 119)$$

$$X_{42} = X_{24} \cup (1, 13, 18, 32, 34, 37, 45, 52, 60, 65, 76, 85, 90, 95, 102, 109, 113, 121)$$

$$X_{46,1} = X_{28,1} \cup (3, 10, 13, 35, 36, 40, 41, 48, 57, 68, 81, 86, 88, 92, 99, 105, 119, 124)$$

$$X_{46,2} = X_{28,2} \cup (1, 13, 21, 27, 33, 39, 45, 53, 58, 69, 75, 82, 90, 99, 106, 107, 113, 119)$$

$$X_{46,3} = X_{28,3} \cup (1, 13, 18, 34, 36, 40, 41, 48, 60, 62, 80, 85, 90, 92, 106, 109, 117, 118)$$

$$X_{46,4} = X_{28,4} \cup (2, 8, 21, 31, 34, 39, 45, 49, 61, 63, 76, 83, 91, 95, 97, 110, 115, 123)$$

$$X_{50} = X_{32} \cup (6, 10, 16, 27, 29, 38, 43, 56, 59, 65, 74, 82, 90, 97, 100, 108, 114, 125)$$

$$X_{56,1} = X_{38,1} \cup (3, 15, 19, 22, 24, 46, 48, 54, 58, 71, 75, 79, 82, 105, 106, 110, 113, 119)$$

$$X_{56,2} = X_{38,2} \cup (4, 10, 18, 22, 30, 41, 48, 56, 63, 67, 74, 80, 84, 98, 101, 108, 112, 125)$$

$$X_{60} = X_{42} \cup 5, 11, 12, 24, 30, 41, 50, 55, 68, 69, 74, 80, 82, 92, 97, 106, 117, 126)$$

6.2. The examples for $b \equiv 0$ (mod 4)

$$X_{12} = (17,20,25,31,40,57,71,76,98,105,121,124)$$

$$X_{16,1} = (2,24,36,37,42,51,61,66,74,78,85,99,103,104,116,126)$$

$$X_{16,2} = (1,20,25,28,43,48,64,68,73,75,77,88,96,104,111,121)$$

$$X_{18} = (1,15,19,32,33,39,45,47,58,63,81,84,88,93,107,111,113,119)$$

$$X_{22} = (6,10,16,19,27,29,39,43,45,47,58,59,65,74,82,90,93,97,108,111,114,125)$$

$$X_{26,1} = (2,7,8,19,21,23,24,29,35,43,47,52,60,65,68,74,80,82,88,92,99,106,108,116,120,124)$$

$$X_{26,2} = (3,10,12,20,21,22,25,28,30,37,45,53,57,59,67,70,74,75,82,98,105,110,118,119,124,125)$$

$$X_{26,3} = (4,5,7,18,26,28,37,46,47,53,58,63,67,74,76,80,84,87,92,94,100,104,110,118,125,126)$$

$$X_{26,4} = (4,10,20,23,27,30,40,44,47,55,58,61,64,72,82,85,90,93,94,108,109,111,116,118,119,124)$$

$$X_{30} = X_{12} \cup (1,15,19,32,33,39,45,47,58,63,81,84,88,93,107,111,113,119)$$

$$X_{34,1} = X_{16,1} \cup (1,15,19,32,33,39,45,47,58,63,81,84,88,93,107,111,113,119)$$

$$X_{34,2} = X_{16,2} \cup (2,10,18,31,34,37,45,52,61,63,79,83,89,95,97,110,112,125)$$

$$X_{36} = X_{18} \cup (6,10,16,27,29,38,43,56,59,65,74,82,90,97,100,108,114,125)$$

$$X_{40} = X_{22} \cup (2, 8, 21, 31, 35, 40, 44, 48, 60, 63, 76, 84, 91, 94, 98, 110, 116, 122)$$

$$X_{44,1} = X_{26,1} \cup (3, 10, 15, 28, 34, 37, 45, 55, 61, 66, 79, 83, 86, 94, 98, 107, 112, 126)$$

$$X_{44,2} = X_{26,2} \cup (1, 15, 19, 31, 33, 38, 46, 48, 58, 64, 80, 84, 88, 94, 106, 111, 112, 120)$$

$$X_{44,3} = X_{26,3} \cup (1, 13, 17, 30, 33, 39, 41, 56, 61, 69, 72, 86, 90, 99, 102, 103, 117, 119)$$

$$X_{44,4} = X_{26,4} \cup (1, 15, 17, 28, 34, 37, 46, 54, 59, 68, 74, 86, 88, 98, 103, 107, 112, 121)$$

$$X_{48} = X_{30} \cup (6, 10, 16, 27, 29, 38, 43, 56, 59, 65, 74, 82, 90, 97, 100, 108, 114, 125)$$

$$X_{52,1} = X_{34,1} \cup (5, 8, 20, 22, 28, 46, 49, 53, 64, 68, 72, 79, 86, 96, 100, 110, 117, 122)$$

$$X_{52,2} = X_{34,2} \cup (4, 8, 14, 30, 33, 38, 46, 53, 60, 70, 78, 82, 87, 92, 99, 103, 116, 126)$$

$$X_{54} = X_{36} \cup (5, 13, 18, 24, 26, 37, 52, 53, 60, 62, 73, 78, 91, 103, 106, 109, 117, 118)$$

$$X_{58} = X_{40} \cup (5, 7, 20, 28, 33, 41, 42, 52, 61, 67, 72, 85, 89, 95, 99, 102, 121, 124)$$

$$X_{62,1} = X_{44,1} \cup (1, 14, 18, 31, 36, 38, 42, 51, 59, 62, 81, 85, 89, 95, 103, 111, 114, 118)$$

$$X_{62,2} = X_{44,2} \cup (5, 13, 18, 27, 35, 40, 42, 44, 60, 66, 73, 78, 87, 92, 100, 107, 122, 126)$$

$$X_{62,3} = X_{44,3} \cup (2, 8, 19, 31, 36, 40, 45, 48, 60, 65, 78, 82, 91, 93, 97, 109, 115, 124)$$

$$X_{62,4} = X_{44,4} \cup (5, 9, 12, 26, 33, 39, 51, 53, 63, 71, 77, 80, 84, 92, 96, 105, 117, 126)$$

6.3. The examples for $b \equiv 3$ (mod 4)

$$X_{15} = (4,11,12,34,46,48,71,77,85,89,94,96,104,116,126)$$

$$X_{21} = (2,8,17,21,23,28,35,40,44,49,61,64,71,73,82,89,94,98,108,121,124)$$

$$X_{25} = (4,11,14,18,23,28,30,41,46,52,53,57,68,69,73,77,83,91,92,99,101,103,109,116,123)$$

$$X_{29} = (2,9,14,17,18,22,26,27,34,38,42,54,56,60,66,69,71,72,73,78,86,96,98,103,106,117,118,122,124)$$

$$X_{33} = X_{15} \cup (1,15,19,32,33,39,45,47,58,63,81,84,88,93,107,111,113,119)$$

$$X_{39} = X_{21} \cup (1,15,19,32,33,39,45,47,58,63,81,84,88,93,107,111,113,119)$$

$$X_{43} = X_{25} \cup (1,15,19,32,33,39,45,47,58,63,81,84,88,93,107,111,113,119)$$

$$X_{47} = X_{29} \cup (1,15,19,32,33,39,45,47,58,63,81,84,88,93,107,111,113,119)$$

$$X_{51} = X_{33} \cup (6,10,16,27,29,38,43,56,59,65,74,82,90,97,100,108,114,125)$$

$$X_{57} = X_{39} \cup (5,12,20,22,24,43,52,53,60,68,72,80,85,102,106,110,117,118)$$

$$X_{61} = X_{43} \cup (6,10,16,27,29,38,43,56,59,65,74,82,90,97,100,108,114,125)$$

6.4. The examples for $b \equiv 1 \pmod 8$

$$X_{17} = (11,12,23,30,31,44,50,57,69,70,76,83,94,110,116,120,123)$$

$$X_{19} = (1,15,19,32,33,39,45,47,58,63,81,84,88,93,107,111,113,119,126)$$

$$X_{27} = (2,5,8,11,18,24,30,34,41,48,50,55,60,62,66,75,79,83,86,91,92,101,106,109,110,112,117)$$

$$X_{35} = X_{17} \cup (11,12,23,30,31,44,50,57,69,70,76,83,94,110,116,120,123)$$

$$X_{37} = X_{19} \cup (2,8,20,25,36,38,49,53,64,66,75,77,91,98,100,102,115,124)$$

$$X_{45} = X_{27} \cup (1,15,19,32,33,39,45,47,58,63,81,84,88,93,107,111,113,119)$$

$$X_{53} = X_{35} \cup (6,10,16,27,29,38,43,56,59,65,74,82,90,97,100,108,114,125)$$

$$X_{55} = X_{37} \cup (6,10,16,22,28,43,48,56,65,67,74,80,82,96,99,110,114,125)$$

$$X_{63} = X_{45} \cup (1,15,19,32,33,39,45,47,58,63,81,84,88,93,107,111,113,119)$$

6.5. The examples for $b \equiv 5 \pmod 8$

$$X_{13,1} = (16,21,22,48,51,65,72,85,99,102,110,117,119)$$

$$X_{13,2} = (6,16,29,42,45,58,74,91,99,100,102,110,114)$$

$$X_{13,3} = (6,10,20,32,41,49,50,66,83,93,101,114,126)$$

$$X_{13,4} = (2,36,37,44,54,59,75,88,90,94,98,102,125)$$

$$X_{23,1} = (1,2,10,27,35,41,46,48,51,68,69,72,73,75,85,89,90,95,96,97,105,107,121)$$

$$X_{23,2} = (7,16,23,29,30,31,39,43,56,58,61,69,71,73,83,84,93,103,106,118,120,121,123)$$

$$X_{23,3} = (1,7,17,19,27,33,34,40,45,57,65,66,70,76,87,90,97,101,105,113,115,119,123)$$

$$X_{23,4} = (2,10,11,21,23,26,31,47,50,59,62,68,73,77,86,88,94,99,107,111,114,119,124)$$

$$X_{31,1} = X_{13,1} \cup (1,14,20,32,33,40,44,47,58,63,81,83,89,93,107,111,114,118)$$

$$X_{31,2} = X_{31,2} \cup (1,13,18,34,36,40,41,48,60,62,80,85,90,92,106,109,117,118)$$

$$X_{31,3} = X_{13,3} \cup (1,15,19,31,33,38,46,48,58,64,80,84,88,94,106,111,112,120)$$

$$X_{31,4} = X_{13,4} \cup 1,14,20,31,32,38,46,49,58,65,79,83,89,95,105,111,112,120)$$

$$X_{41,1} = X_{23,1} \cup (3,7,20,25,34,43,47,52,65,67,74,76,91,93,100,104,120,122)$$

$$X_{41,2} = X_{23,1} \cup (2,9,17,33,35,37,46,50,60,65,80,82,90,94,96,108,114,125)$$

$$X_{41,3} = X_{23,3} \cup (2,10,20,31,32,38,43,52,61,62,77,85,89,95,99,111,112,124)$$

$$X_{41,4} = X_{23,4} \cup (1,13,18,32,34,37,45,52,60,65,76,85,90,95,102,109,113,121)$$

$$X_{49,1} = X_{31,1} \cup (3,10,15,24,35,37,50,55,64,66,78,79,86,95,101,103,112,126)$$

$$X_{49,2} = X_{31,2} \cup (3,10,15,24,35,37,50,55,64,66,78,79,86,95,101,103,112,126)$$

$$X_{49,3} = X_{31,3} \cup (4,8,21,27,30,39,45,53,60,63,72,87,89,98,99,110,116,122)$$

$$X_{49,4} = X_{31,4} \cup (5,13,18,22,26,41,48,56,60,66,73,78,87,104,106,108,113,121)$$

$$X_{59,1} = X_{41,1} \cup (5,8,13,30,33,39,45,53,59,71,78,83,86,92,99,102,117,126)$$

$$X_{59,2} = X_{41,2} \cup (3,10,15,26,34,38,47,55,64,66,78,79,86,92,100,107,113,126)$$

$$X_{59,3} = X_{41,3} \cup (4,11,12,25,30,42,50,54,68,69,73,81,83,92,96,107,116,126)$$

$$X_{59,4} = X_{41,4} \cup (8,14,20,22,24,36,49,53,58,64,70,81,82,104,108,115,116,120)$$

7. DISCUSSION

This paper studies the non-isomorphic $E(s^2)$-optimal supersaturated designs in ssd(m,N). In particular, for $N = 10$ runs, we give all the possible minimal row Gram structures. As we have seen, for $N = 10$ all minimal row Gram structures are realizable by some supersaturated design. However, there is no clear pattern as to how many non-isomorphic $E(s^2)$-optimal supersaturated designs exists for a given number of factors m. Nevertheless, it is interesting to observe that for $b \equiv 1 \pmod 8$ and $b \equiv 3 \pmod 8$, we found a unique $E(s^2)$-optimal supersaturated design. For the remaining cases, the number of non-isomorphic $E(s^2)$-optimal supersaturated designs range from 1 to 4.

This leaves us with some questions for further inquiry:

1. Are additional, statistically meaningful criteria that differentiate when more than one non-isomorphic $E(s^2)$-optimal supersaturated design exists?

2. Are all possible minimal row Gram structures realizable for $N = 14$? For general $N \equiv 2 \bmod 4$?

3. What about the case $N \equiv 0 \bmod 4$?

4. In many cases, Section 3 describes the minimal row Gram structures via graphs. Is there a connection between the properties of those graphs and the properties of the $E(s^2)$-optimal designs.

References

Booth, K.H.V. and Cox, D.R. (1962). Some systematic supersaturated designs. In *Technometrics*, 4: 489–495.

Bulutoglu, D.A. and Cheng, C.S. (2004). Construction of $E(s^2)$-optimal supersaturated designs. In *The Annals of Statistics*, 32: 1662–1678.

Butler, N.A., Mead, R., Eskridge, K.M., and Gilmour, S.G. (2001). A general method of constructing $E(s^2)$-optimal supersaturated designs. In *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63: 621–632.

Cheng, C.S. (1997). $E(s^2)$-optimal supersaturated designs. In *Statistica Sinica*, 7: 929–939.

Das, A., Dey, A., Chan, L.Y., and Chatterjee, K. (2008). On E(s2)-optimal supersaturated designs. In *Journal of statistical planning and inference*, 138 (12): 3749–3757.

Domagni, K.F., Hedayat, A.S., and Sinha, B.K. (2021). On the nature of saturated 2^k-factorial designs for unbiased estimation of non-negligible parameters. In *Journal of Statistical Planning and Inference*, 211: 33–40.

Domagni, K.F. and Neubauer, M.G. (2025). Online appendix. `http://www.csun.edu/~hcmth007/ssdDN.pdf`. Online appendix.

Eskridge, K.M., Gilmour, S.G., Mead, R., Butler, N.A., and Travnicek, D.A. (2004). Large supersaturated designs. In *Journal of Statistical Computation and Simulation*, 74 (7): 525–542.

Gale, D. (1957). A theorem on flows in networks. In *Pacific Journal of Mathematics*, 7 (2): 1073–1082.

Jones, B. and Majumdar, D. (2014). Optimal supersaturated designs. In *Journal of the American Statistical Association*, 109 (508): 1592–1600.

Nguyen, N.K. (1996). An algorithmic approach to constructing supersaturated designs. In *Technometrics*, 38: 69–73.

Ryser, H.J. (1957). Combinatorial properties of matrices of zeros and ones. In *Canadian Journal of Mathematics*, 9: 371–377.

Singh, R. and Stufken, J. (2023). Selection of two-level supersaturated designs for main effects models. In *Technometrics*, 65 (1): 96–104.