FINAL PROGRAM REPORT

High Dimensional Inference and Random Matrices 2006-2007

1. Program and its Objectives:

Random matrix theory lies at the confluence of several areas of mathematics, especially number theory, combinatorics, dynamical systems, diffusion processes, probability and statistics. At the same time, random matrix theory may hold the key to solving critical problems for a broad range of complex systems from biophysics to quantum chaos to signals and communication theory to machine learning to finance to geophysical modeling. This Program was a unique opportunity to explore the interplay of stochastic and mathematical aspects of random matrix theory and its applications.

The aim of the program was to bring together researchers interested in the theory and applications of random matrices to share their results, discuss new research directions and develop collaborations. The program concentrated on large-dimensional random matrices and the problems that make use of them. In particular, emphasis was on how developments in random matrix theory might impact statistical inference in high dimensional systems.

The program has had two parts. In the fall, there was an emphasis on statistical inference and the theory of random matrices. In the spring semester, the focus shifted to issues arising in connection with geometry and random matrices.

2 Core Group

A core group of researchers gathered at SAMSI during the Fall Semester, 2006 and this group was complemented by a group on the West Coast that gathered every Friday and held a video-conference with SAMSI. This link facilitated significant exchange between the two groups. In addition, various people joined the program through Webex connections to the working groups.

2.1 *Senior researchers (at SAMSI for significant periods of time in Fall, 2006):*

- Jianqing Fan (Princeton University)
- Eitan Greenshtein (Purdue University)
- Thomas Guhr (Lund University)
- Christian Houdre (Georgia Institute of Technology)
- Helene Massam (York University)
- Peter Miller (University of Michigan)
- Greg Rempala (University of Louisville)
- Don Richards (Penn State University)
- Nan Wermuth (Chalmers U of Technology)
- Ofer Zeitouni (University of Minnesota)

2.2 *Senior Researchers (Spring, 2007)*

- Mikhail Belkin (Ohio State University)
- Yoonkyung Lee (Ohio State University)
- Feng Liang (University of Illinois Urbana Champaign)

2.3 *Key Berkeley node researchers*

- Noureddine El Karoui (University of California Berkeley)
- Iain Johnstone (Stanford University)
- Peter Bickel (University of California Berkeley)
- Craig Tracy (University of California Davis)
- Bin Yu (University of California Berkeley)

2.4 *New researchers*

- Debhashis Paul (University of California Davis)
- Makram Talih (Hunter College)
- Ming Yuan (Georgia Institute of Technology)

2.5 *Postdoctoral fellows*

- Manjunath Krishnapur
- Jayanta Pal
- Bala Rajaratnam
- Cari Kaufman (Joint with Complex Computer Models Program)
- Elain Spiller (Joint with Complex Computer Models Program)

2.6 *Local faculty*

- Ilse Ipsen (North Carolina State University)
- Yufeng Liu (University of North Carolina)
- Sayan Mukherjee (Duke University) (Spring)
- Jack Silverstein (North Carolina State University)
- Len Stefanski (North Carolina State University)
- Young Truong (University of North Carolina)
- Stephanos Venakides (Duke University)
- Li Lexin (North Carolina State University) (no teaching release)
- Haipeng Shen (University of North Carolina) (Spring)

2.7 *Graduate students*

- Sergei Belov (Duke University)
- Hongyan Cao (University of North Carolina)
- Zhenglei Gao (Duke University)
- William Lefew (Duke University)
- Trevia Litherland (Georgia Institute of Technology)
- Jinchi Lv (Princeton University)

- Xingye Qiao (University of North Carolina)
- Teresa Selee (North Carolina State University)
- Dhruv Sharma (North Carolina State University)
- Hua Xu (Georgia Institute of Technology)
- Yingying Fan (Princeton University)
- Yufan Zhao (University of North Carolina)

3. Program Organization

3.1 *Opening workshop*

The Opening Workshop for the SAMSI program on High Dimensional Inference and Random Matrices was held Sunday-Wednesday, September 17-20, 2006, at the Radisson Hotel RTP in Research Triangle Park, NC. It was preceded, on Sunday, September 17, with tutorials by Craig Tracy, and Ofer Zeitouni.

The goal of the opening workshop was to allow community input to formation of the working groups for the program, as well as promote engagement (via web, teleconference, and videoconference) of those who will not be resident at SAMSI during the program. The workshop program focusses heavily on open problems in high dimensional inference and random matrix theory for which solutions are not currently available.

The workshop was organized by Iain Johnstone, Peter Bickel, Hélène Massam, Douglas Nychka, and Craig Tracy.

The workshop included a number of distinguished speakers. Apart from the Distinguished Lecture of David Donoho (see below), featured were also overview lectures by Roland Speicher and Alan Edelman.

During the Tuesday evening, the program leaders committee met with some of the key participants who were to be present at SAMSI. A first cut at the working groups was made. After discussions the following day, a strong list of working groups was formed and the participants signed up for what interested them. There was an extraordinary response to the working group call, with almost all of the workshop participants staying around for the working group formation.

3.2 *Bayesian Focus Week*

The Bayesian Focus Week workshop: October 30th to November 3rd, 2006 at the Radisson Hotel RTP in Research Triangle Park, NC.

The themes of the workshop were inference in high-dimensional graphical models, choice of priors for efficient model search in Gaussian and log-linear models, and algorithms for model selection.

Problems of inference in Gaussian graphical models include high dimensional covariance estimation, estimation of eigenvalues, computation of moments, and Gaussian graphical models for time series. Theoretical as well as implementation properties of various priors will be considered for model selection.

The emphasis in the workshop was on discussions, interaction, problem solving and the identification of new problems in a cross-disciplinary setting, concentrating on high-dimensional problems and exploring the possible use and applications of large random matrices.

The workshop was organized by Hélène Massam, Peter Bickel, and Mike West.

3.3 *Large Graphical Models and Random Matrices Workshop*

November 9-11, 2006 at the MCNC Auditorium and the Radisson Hotel RTP in Research Triangle Park, NC. The workshop focused on the following topics: transforming real-valued matrices to study induced associations in linear systems and transforming binary matrices to study the preservation of independencies in Markov graphs, matrix and path criteria for Markov equivalence for separation in Markov graphs and for identification of corresponding models under given distributional assumptions, the integration of clustering and of censoring into graphical models, issues of model fitting, model selection and model uncertainty for graphical models, and consequences for a given generating process of omitting variables and of conditioning on variables or on the propensity score

The workshop was organized by Nanny Wermuth, Hélène Massam, and David Cox.

3.4 *Workshops with NCAR*

A number of workshops with NCAR were planned. Some of these have been primarily organized through the Complex Computer Models Program but involve some HDRM participants (for instance, the Geophysical Models Workshop, Nov 13-14.)

A workshop is planned for May, to be held at NCAR. This workshop will emphasize the problems in random matrix theory that arise in large geophysical models. The workshop will be entitled: Application of Random Matrices Theory and Methods, be held May 7-9 At NCAR, Boulder, Colorado and organized by Thomas Bengtsson, MontseFuentes, and Peter Bickel.

3.5 *Geometry workshop*

Geometry, Random matrices, and Statistical Inference was held January 16-19, 2007 at the NISS Building in Research Triangle Park, NC.

The four day workshop kicked off the semester long focus on Geometry and Random Matrices. Both algorithms and the fundamental mathematical objects computed by the

algorithms were stressed. This workshop was followed by a semester long working group on "Geometry, Random Matrices, and Statistical Inference."

The format of the workshop had two talks in the morning and one in the afternoon. All talks were one hour leaving considerable time for discussion between.

The workshop was organized by Misha Belkin (The Ohio State University, Chair), Feng Liang (University of Illinois at Urbana-Champaign), and Sayan Mukherjee (Local Scientific Coordinator).

3.6 *Transition workshop*

The transition workshop was held at the American Institute of Mathematics in Palo Alto: April 10-13, 2007 as an ARCC workshop. Apart from the very pleasing bracketing of the program as having been both initiated and culminated at AIM, the nature of this transition workshop was ideally suited to the format of an ARCC workshop. Indeed, the leadership of the ARCC team in facilitating the group to collect its thoughts and focus their ideas into a blueprint for future research was beneficial and efficient for achieving the goals of the workshop.

This workshop brought together a group of applied mathematicians active in random matrix theory with theoretical statisticians (and probabilists) concerned with high dimensional inference particularly via eigen-structure methods. It also engaged methodologically oriented researchers from application domains, such as climate prediction, in which large p data analysis has long played a major role.

The format of the meeting was that during the first day and a half, reports on the working group activities from the fall SAMSI program were delivered. Each of the four main reports was given by a duo of one senior and one junior researcher:

1. Bayesian Methods and Graphical Models: Helene Massam and Bala Rajranatham
2. Geometric Methods: Makram Talih and Armin Schwatzmann
3. Geometry and Statistical Inference: Sayan Mukherjee and Misha Belkin
4. Regularization and Covariance: Peter Bickel and Debashis Paul
5. Multivariate Distributions: Don Richards and Nourredine El Karoui.

In addition, a talk on the results of the wireless communications collaboration was given by Jack Silverstein. The emphasis was on discussions about achievements of the working groups and future directions.
On Thursday morning, there were three short unplanned presentations from participants who had something they wanted to communicate to the others. Nanny Wermuth spoke on matrix representations and independencies in directed acyclic graphs. Jamie Mingo gave his view on the computation of moments of random matrices with combinatorial arguments in the framework of free probabilities. Boaz Nadler spoke on eigenvalue estimation and stochastic perturbation theory. These three speakers were all participants in the Opening SAMSI workshop who had not been able to participate in the working groups and they all gave talks that presented a different, original point of view on classical problems in the area of large random matrices. In the afternoon, an attempt was

made to impose four or five working groups as was done the day before but there was "spontaneous disobedience". The active participants had evolved beyond the common discussion stage. They now wanted to work in small groups of two or three people on precise topics and that is exactly what happened.
On Friday morning, Mylene Maida and Eitan Greenshtein gave short presentations on some of their current problems and these were followed by many questions.

3.7 *Distinguished Lectures*

Two SAMSI Distinguished Lectures were held as part of the program. Each was part of a workshop.

David Donoho, Stanford University, Monday September 18, 2006
The Breakdown Point of Model Selection When There Are More Variables Than Observations

David Cox, Nuffield College, Oxford, Thursday November 9, 2006
Some Statistical Challenges Arising from an Issue in Veterinary Epidemiology

4. Activities

4.1 *Working groups*

The working groups met regularly throughout the program to pursue particular research topics identified in the kickoff workshop.

- Climate and Weather
- Wireless Communications
- Universality
- Regularization and Covariance
- Geometric Methods
- Multivariate Distributions
- Graphical Models/Bayesian Methods
- Estimating functionals of a high dimensional sparse vector of means

4.2 *Berkeley node*

Each Friday during the fall semester, a video-conference link was established between SAMSI and UC Berkeley. This allowed a large number or participants to interact regularly with the SAMSI working groups. Participation in the video link-up was sustained and impressive often with 30 people connected

4.3 *Courses*

4.3.1 *Fall 2006 SAMSI course on Random Matrices*

This course was an introduction to properties of eigenvalues of classes of random matrices fundamental to several areas of application, including multivariate statistics, high energy physics, numerical analysis, and wireless communications. Most of the results are expressed in terms of limit theorems, as the dimensions of the matrices increase, the significance of which enables the understanding of spectral behavior for random matrices of large dimension. The main results covered include the limiting behavior of empirical measure of the eigenvalues (law of large numbers and CLT's for linear statistics), extreme eigenvalues, distribution of largest eigenvalue, concentration inequalities and large deviation theory for the empirical measure and maximal eigenvalue.

The basic mathematical tools used to prove these results were introduced, among them being: moment methods, Stieltjes transforms, concentration inequalities and large deviations, and integrable systems. Lectures on applications to multivariate statistics were also included.

The course was team-taught by individuals who are renowned experts in random matrices.

Schedule of Instructors and topics covered:

- September 7,14 Jack Silverstein and Bala Rajaratnam: Introduction to random matrices and the rmtools software.
- September 21,28 October 5,12 Ofer Zeitouni: Method of moments for law of large numbers, CLT's and largest eigenvalue; concentration; GOE-GUE; large deviations; Fredholm determinants, determinantal point processes.
- October 19,26 Jack Silverstein: Stieltjes transform methods.
- November 2,9 James Mingo: Free probability and random matrices
- November 16,30 Peter Miller: Asymptotics for general orthogonal polynomials; Riemann-Hilbert approach.
- December 7,14 Debashis Paul: Multivariate statistical applications

4.3.2 *Spring 2007 SAMSI course on Geometry, Random Matrices, and Statistical Inference*

From the perspective of inference, clustering, and machine learning, geometric ideas have been gaining greater emphasis. One reason for this has been the realization that predictive models with a small amount of labeled data can be greatly improved by incorporating unlabeled data. Thus the geometry of the marginal distribution provides salient and compelling information in many real world problems.

This insight has led to a variety of statistical models and algorithms as well as the study of a variety of mathematical objects. A non-exhaustive list follows: spectral clustering, nonlinear dimensionality reduction, manifold learning, learning homologies, topological persistence, semi-supervised learning, non-parametric semi-supervised Bayesian models, the Laplace-Beltrami operator, graph diffusion models on manifolds, random projections. Most if not all of the above topics are intimately related to the study of random matrices

either from an algorithmic perspective or from the perspective that the structure of a random matrix depending on data drawn from a measure is fundamental in understanding the topic. These topics are being presented from the perspective of Statisticians, Computer Scientists, and Mathematicians.

The course started with a lightning review of statistical inference, topology, and differential geometry and then proceeded to seminar format. There will be a final project consisting of any of the following:

1. paper review
2. algorithm/model development
3. data analysis
4. theoretical analysis

Instructors: Misha Belkin, Sayan Mukherjee, and Yury Mileyko

5. Working Group Reports

5.1 *Climate and Weather*

Group Leaders: Serge Guillas, Cari Kaufman, Doug Nychka, Debashis Paul
Webmasters: Cari Kaufman and Elaine Spiller

The climate and weather group met as a reading group in preparation for the joint SAMSI/NCAR workshop held in November, at which point the climate and weather groups from the Computer Models and Random Matrices programs merged. The group focused primarily on principal component analysis and the ways it is used by climate scientists. In particular, we studied methods for detection and attribution of climate change. We also had two guest speakers, Chunsheng Ma, who spoke on space-time covariance functions that might be appropriate for climate data, and Andrew Gettelman, who gave a remote tutorial from NCAR on climate modeling.

The group developed several ideas for research directions.

1) The usual tests and asymptotic distributions in principal component analysis assume the rows of the data matrix are exchangeable, which clearly does not hold in the case of climate observations that are taken over time. How should one choose the number of components to retain in this case? Jonty Rougier and a graduate student are working on a permutation test.
2) Can better estimates of the covariance matrix be used in the "fingerprint" methods for climate change detection? Current practice is to use the sample covariance matrix.

3) Can the covariance functions described by Chunsheng Ma (2005), which are negative over certain distances, be used to capture large-scale oscillation patterns in the climate system? In particular, can fitting these models to climate models and observations separately be used as a diagnostic tool for assessing how well the climate model captures

such patterns? Cari Kaufman and Chunsheng Ma explored this question for some pressure data.

4) An alternative to principal component analysis proposed by Debashis Paul is to first premultiply the data by the square root of a weighting matrix W, which would concentrate the resulting eigenvectors in certain spatial regions.

5.2 *Wireless Communications*

Group Leader: Jack Silverstein

In this group, we looked at several problems involving random matrices arising in the context of wireless communications. J. Silverstein and D. Paul have been involved in a project that looks into the behavior of the eigenvalue statistics, such as empirical spectral distribution, separation of eigenvalues from the bulk spectrum etc for the random matrices appearing in MIMO and CDMA systems. The problems considered are also relevant in spatio-temporal signal processing. In the latter case there is also interest in investigating the properties of the sample eigenspaces.

5.3 *Universality*

Group Leader: Peter Miller
Webmaster: Manjunath Krishnapur

The Universality Working Group nucleated at the Opening Workshop held September 17 - 20, 2006. Three topics emerged between the Opening Workshop and the first meeting of the group that were of most interest to participants: (i) Universality of Circular Ensembles (random unitary matrices), (ii) "Beta-ensembles" (generalized eigenvalue statistics beyond the basic "threefold way" of matrix symmetries; representation and sampling issues), and (iii) Differential operator analogues of random matrix theory. The first order of business for the Working Group was to have a focused discussion of these three topics with the aim of narrowing our scope to a list of specific questions within each topic.

The Universality Working Group functioned primarily as a venue for informal presentations by and for the group participants. All of the three topics listed in the paragraph above were addressed by these presentations and the ensuing discussions. Speakers included M. Huber ("Some experiments on tridiagonal matrices for beta ensembles"), M. Krishnapur ("Circular ensembles", "An invariance principle with applications to random matrix theory and last passage percolation", and "Non-Hermitian random matrices and the circular law"), P. Miller ("Asymptotics of orthogonal polynomials on the unit circle by Riemann-Hilbert techniques", "D-bar methods for orthogonal polynomials", and "Zakharov-Shabat eigenvalue problems"), D. Paul ("Beta ensembles and tridiagonal matrix models"), and S. Venakides ("Zakharov-Shabat operators as analogues of random matrix models" and "Riemann-Hilbert techniques"). Notes from all of these presentations can be found on the SAMSI website.

The Universality Working Group also had a significant interaction with the "Berkeley Node" of SAMSI. Via the videoconference link the Working Group participated "live" in presentations made in Berkeley by L. Choup, A. Dembo, S. Peche, and A. Soshnikov. This interaction was very useful for both the local participants and those in the studio in Berkeley.

Among the most regular local participants in the group were S. Belov (Duke), M. Huber (Duke), M. Krishnapur (UNC/SAMSI), W. LeFew (Duke), P. Miller (Michigan/UNC/SAMSI), D. Paul (UC Davis/SAMSI), E. Spiller (Duke/SAMSI), and S. Venakides (Duke). Regular remote participants included F. Goetze (Bielefeld, Germany) and I. Johnstone (Stanford). A substantial fraction of the regular participants were graduate students and postdocs; this suggests that the Universality Working Group had a positive educational component.

5.4 *Regularization and Covariance*

Group Leaders: Peter Bickel, Eitan Greenshtein, Debashis Paul, Noureddine El Karoui
Webmasters: Debashis Paul and Noureddine El Karoui

The general focus for this working group was regularization methods in high dimensional statistical inference with emphasis on estimation of high dimensional covariance matrices. Discussions on the general notion of regularization were centered mainly around the papers of Bickel and Li (2006) and of Breiman (1996). Two key themes of the discussions were: (i) regularization as means of obtaining better prediction performance in high dimensional data analysis problems, and (ii) regularization as a way of selecting the correct model in a parametric statistical framework.

Unlike in classical estimation, where the minimization of squared error loss is a widely accepted procedure, many loss functions are popular in the context of estimation of the population covariance. Examples are, Frobenius norm, Kullback-Liebler distance under Gaussian assumption, the spectral norm, and many more. Of course in choosing a criterion or a loss, mathematical convenience is an important issue, but a better understanding of the appropriate loss for a specific context may be desirable. In applications of regularization techniques to covariance estimation, we had discussions about estimation under structural assumptions on the population covariance matrix, and also in a nonstructural context.

Nonstructured context: Given an observed random matrix Q, with $\Sigma = E(Q)$, in many statistical problems the goal is to find a vector β with largest (or smallest) value of the quadratic form $\beta^T\Sigma\beta$, where the selection of the vector β is based on the observed Q. When Q_{pxp} is a Wishart matrix, and $p \gg n$, based on a sample of size n, there is no hope of finding a vector β, based on Q, that even nearly the minimizes/maximizes $\beta^T\Sigma\beta$. This is true unless we have strong assumptions on the structure of Σ. When there are no such assumptions, a reasonable approach is to limit ourselves to a subset B of vectors β satisfying certain constraints. The goal is to find nearly the best vector β in B. Of course

the larger B is, the better. This has to do with the concept of persistence, as described in Greenshtein and Ritov (2004) (see also the tutorial by Greenshtein).

A popular choice of B is to take it as an l_1 ball with an appropriate radius. Alternatively introduce an l_1 constraint on β as in Lasso (Tibshirani (1996)). In practice one minimizes $\beta^T Q\beta$, subject to the l_1 constraint. Zou and Hastie (2005) proposed to consider a combination of constraints, e.g., l_1 and l_2 constraints through a procedure they termed '*elastic net'*. A reasonable combination of constraints, proposed by Greenshtein, seems to be an l_1 constraint, combined with a constraint on the estimated quantity $\hat{W}ar(\beta^T Q\beta)$ of $Var(\beta^T Q\beta)$. Note that the last constraint is random. It is hoped that this additional constraint will lead to an improvement in the performance of the estimator.

Yuan and Lin (2005), considered estimation of a non-structured covariance matrix using the method of maximum likelihood under an l_1 constraint on the off-diagonal entries of the covariance matrix (see also the tutorial by Yuan). This involves a sophisticated convex optimization method. However, their set up is that of $p < n$. Peter Bickel, during one of the Berkeley node seminars, referred to similar work (with a different penalty) being carried out by Levina and Zhu. This avenue and the study of the statistical properties of the estimators obtained by such penalized empirical risk minimization procedures may be a fruitful area of research, particularly in the situation when p and n are roughly of the same size.

Structured context: We discussed several different structural assumptions about the covariance matrix that are applicable in various problems. One central question was how to translate available information about the problem into the specification of a structure for the covariance. One natural structure is that of a sparse covariance matrix. One subclass of this consists of matrices that, under an appropriate permutation of variables, have a *block diagonal structure*. Some discussion on the relevance of this in certain econometric problems may be found in Guhr and Kaebler (2003) (see also tutorial by Guhr), and in a tutorial by Paul. It was suggested that estimation of maximal eigenvectors could be helpful in identifying such a structure. Similar methods have been used to deal with the problem of *graph partitioning*. Some discussions involved the more general setting of graphical models and of finding the zeroes of the *precision matrix* Σ^{-1}. In this context we discussed the paper by Meinshausen and Bühlman (2006). See also the aforementioned paper by Yuan and Lin.

Another structure that has been studied and discussed is that of a factor model. See the tutorial by Lv and the manuscript by Fan *et al*. (2006) for more details. Estimation of Σ and Σ^{-1} is studied in the last article. It also includes some discussions on the choice of loss functions. A similar model has been used in the context *of functional data analysis* (see also the tutorial by Paul), and recent results from random matrix theory (c.f. Paul (2006)) shed light on the statistical properties of the estimates obtained under the model.

A covariance matrix of a Toeplitz form is natural in various statistical contexts, e.g. involving a stationary time series. In Bickel and Levina (2006), estimation of a covariance matrix Σ is studied, under the assumption that $|\sigma_{ij}|$ converges to zero as $|i-j|$

approaches infinity, in such a way that the contribution of the elements away from the diagonal of Σ become negligible. Here $\Sigma = \Sigma_{\infty x \infty}$. A special case of this involves matrices that are certain additive perturbations of Toeplitz matrices. See also the tutorial by Bickel. The suggested estimation method involves banding, and the loss function is the spectral norm of the difference. Partly motivated by this approach, Zeitouni and Anderson studied the problem of banding of certain classes of random hermitian matrices in a general context and derived new asymptotic results on the behavior of the empirical spectral measure of such matrices (see the lecture notes by Zeitouni).

Various classical statistical procedures were proposed regarding the estimation of the mean vector and covariance matrix in high-dimensional context. As a part of Berkeley Node seminar, Eitan Greenshtein presented a method as an alternative to the Empirical Bayes procedure for estimating the mean vector in Gaussian signal-plus-noise model. The method has excellent empirical performance even when the true signal is dense. Whether one can design an estimation procedure for the covariance that can perform under a wide range of sparsity of the entries of the matrix is an open question. Also, Martin Wainwright gave a presentation in the Berkeley Node seminar about the model selection properties of L_1 regularization in the context of Gaussian linear regression.

On the application side, Young Truong gave an insightful lecture on fMRI data and the potential regularization procedures that can be utilized to answer some of the questions.

In addition to these activities, members of this working group actively participated in collaborative activities with the members of other working groups. We give a brief description of these activities below.

- *Interaction with Climate group:* Problems in climatology and atmospheric science involve dealing with large data on processes that vary in both space and time. Computation of large covariance matrices is a part of many well-known methods (e.g. Kalman filtering) that are used in these contexts. Our interaction with the Climate group involved learning about the problems arising from these fields and assessing the possibility of developing statistical methodologies for dealing with the questions that arise. Carie Kaufman gave a talk on the use of *tapering* for an efficient computation of estimates of parameters describing a spatial process. There is clearly a lot of scope of collaborative work with scientists working in this area.

- *Interaction with Universality group:* Many inference questions relating to large dimensional covariance matrices can be addressed by using tools from random matrix theory. The interaction with Universality group provided an opportunity to know some of the latest developments in this fast-growing field.

- *Interaction with Graphical models group:* We discussed several applications in the context of structured covariance matrices that can be formulated in the framework of Gaussian graphical models. Some model selection issues in this context were addressed through interactions between our group and the Graphical models group.

References

1. Bickel, P., and Levina, E. (2006): Regularized estimation of large covariance matrices. *Technical report*.

2. Bickel, P., and Li, B. (2006): Regularization in Statistics. *Test*, **15**, 271-344.

3. Breiman, L. (1996): Heuristics of instability and stabilization in model selection. *Annals of Statistics*, **24**, 2350-2383.

4. Fan, J., Fan, Y., and Lv, J. (2006): High dimensional covariance matrix estimation using a factor model. *Technical report*.

5. Greenshtein, E., and Ritov, Y. (2004). Persistence in high dimensional linear predictorselection and the virtue of over parametrization. *Bernoulli,* **10**, 971-988.

6. Guhr, T., and Kaebler, B. (2003): A new method to estimate the noise in financial correlation matrix. *Journal of Physics A: Math. Gen*., **36**, 3009-3032.

7. Meinshausen, N., and Bühlman, P. (2006): High dimensional graphs and variable selection with Lasso. *Annals of Statistics*, **34**, 1436-1462.

8. Paul, D. (2006): Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, (to appear).

9. Tibshirani, R. (1996): Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, Series B*, **58**, 267-288.

10. Yuan, M., and Lin, Y. (2005): Model selection and estimation in the Gaussian graphical model. *Biometrika*, (to appear).

11. Zou, H., and Hastie, T. (2005): Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society, Series B*, **100**, 301-320.

5.5 Geometric Methods

Group Leader: Makram Talih
Webmaster: Jayanta Pal

After the initial organizational meeting, participants in the Geometric Methods working group decided to focus on building an intuitive understanding of the geometry of the cone of symmetric positive definite matrices, with a view to exploit their intrinsic structure for modeling and inference in highdimensional covariance matrices. To achieve this goal, the group discussed two important papers in the field: one by Moakher (2005) on defining the geometric mean of symetric positive-definite matrices using the notion of geodesic

curves and on building a gradient-descent algorithm for its numerical approximation; the second by Fletcher & Joshi (2004), who, in the context of diffusion-tensor magnetic resonance imaging, introduce principal component analysis, coined principal geodesic analysis, whose modes of variation are geodesic lines in the cone of diffusion tensors (ie. 3X3 variance-covariance matrices).

The next order of business for the group, and one of the group's stated main themes of research, has been to investigate probability distributions on covariance manifolds, in particular, the cone of symmetric positive definite matrices, which is looked upon as a Riemannian manifold. We were especially interested in reading about some recent findings by a research group at the INRIA, based in Sophia-Antipolis, France, especially Lenglet et al (2004) and Pennec (2004). Working group member Dr. Armin Schwartzman was keenly active in this endeavor, and gave an informal talk about his own research on distributions for positive definite matrices on 11/30/06. He was also later invited to present his findings in the follow-up workshop on Geometry, Random Matrices, and Statistical Inference, held at SAMSI in January 2007. The study of probability distributions on the cone of positive definite matrices is intimately related to the question of defining a Normal distribution thereon, since the latter is the principal paradigm for such distributions, but also since it appears as the limiting distribution in the central limit theorem. Thus, the group highly benefited from close interaction with another of the HDRM working groups, namely the Multivariate Distributions working group, with which most group members were also affiliated. Prof. Don Richards, especially, was instrumental in leading the discussion on 11/30/06 about the important role of the Helgason-Fourier transform in the derivation of a central limit theorem on spaces of positive definite matrices. On 12/14/2006, Prof. Richards also gave a presentation on minimax estimation for the deconvolution problem over the space of positive definite matrices.

The theme of inference for time-varying structures has been mainly taken up by the group's leader, Dr. Makram Talih, who has presented work on constructing a Markov chain on the cone of positive definite matrices such that the path between consecutive points is a geodesic segment. Dr. Talih's line of research is motivated by the problem of estimating the covariance structure in the multivariate Normal model when the underlying matrix is driven by a hidden Markov chain. Dr. Talih's work has greatly benefited from the working group's input and suggestions, and was featured in the Bayesian focus week in October 2006, as well as in the SAMSI Education and Outreach program's two-day undergraduate workshop in November 2006.

Motivated by the HDRM program's regularization theme, the geometric methods working group has also been discussing the use of geodesic distance between variance-covariance matrices in defining sensible loss functions that are invariant to inversion and orthogonal transformations. In this respect, the geometric perspective provides valuable insight for the study of shrinkage and regularization.

5.6 *Multivariate Distributions*

Group Leaders: Don Richards, Iain Johnstone
Webmaster: Jayanta Pal

The group leaders, Iain Johnstone and Don Richards, with assistance from Jayanta Pal, organized the group activities by finding speakers for each weekly meeting, coordinating group activities with the SAMSI staff and directorate, and maintaining contact between group members located at SAMSI or elsewhere.

The webmaster, Jayanta Pal, maintained the group's web page at <http://www.samsi.info/200607/ranmat/workinggroup/md/ranmat-md.html>. The web page was updated at least twice weekly, listing upcoming group activities, talks and discussions, papers and smartboard notes, and interactions with other working groups. A key feature of the group meetings was the participation of off-site members through the WebEx system; this enabled high virtual attendance by off-site group members. The website has been maintained into this semester and contains all notes and papers which formed the basis for group discussions and activities.

Throughout the semester, there also was strong interaction with the Working Group on Geometric Methods. Makram Talih, leader of the Geometric Methods group, was particularly helpful in fostering this joint interaction. As a consequence, the two groups held three joint weekly meetings. At other meetings of the Geometric Methods group, members of the Multivariate Distributions group offered their perspectives on discussions ranging from Central Limit Theorems on spaces of positive definite matrices to concepts of geometric means for positive matrices.

Members of the Working Group on Multivariate Distributions and their affiliations are as follows: Jeongyoun Ahn (University of Georgia), Edo Airoldi (Carnegie-Mellon University), Leonard Choup (UC-Davis), Noureddine El Karoui (UC-Berkeley), Friedrich Goetze (Bielefeld University), Thomas Guhr (Lund Inst. of Technology), Iain Johnstone (Stanford University), Plamen Koev (MIT), Stas Kolenikov (University of Missouri), Yoshihiko Konno (Japan Women's University), Helene Massam (York University), Jing Naihuan (NC State University), Jayanta Pal (SAMSI), Debashis Paul (UC-Davis), Xingye Qiao (UNC), Bala Rajaratnam (SAMSI), Don Richards (Penn State University), Igor Rumanov (UC-Davis), Dhruv Sharma (NC State University), Leonard Stefanski (NC State University), Ming Yuan (Georgia Tech).

The weekly meeting activities of the group were as follows:

Date: 09/28/06
Speaker: Bala Rajaratnam
Topic: Marginal likelihood for the eigenvalues of covariance matrices

Date: 10/05/06
Speaker: Donald Richards
Topic: Multimodality of the likelihood function for the Behrens-Fisher problem

Date: 10/12/06
Speaker: Donald Richards
Topic: An introduction to zonal polynomials and hypergeometric functions of matrix argument

Date: 10/19/06
Speaker: Plamen Koev
Topic: The combinatorial definition of the Schur, zonal, and Jack polynomials

Date: 10/26/06
Speaker: Donald Richards
Topic: Generalizations of the Wishart distribution arising from monotone incomplete multivariate normal data

Date: 10/30/06 - 11/3/2006
The group members attended the SAMSI program, "Bayesian Focus Week"

Date: 11/9/06 - 11/11/2006
The group members attended the SAMSI program on "Large Graphical Models and Random Matrices"

Date: 11/16/06
Speakers: Jayanta Pal and Donald Richards
Topic: Discussion on MIMO capacities, representation theory and eigenvalue computations.

Date: 11/30/06
Speaker: Noureddine El-Karoui
Topic: Finite point processes and eigenvalues of random matrices

Date: 12/07/06 (Joint meeting with the working group on Geometric Methods)
Speaker: Hongtu Zhu, UNC Chapel Hill
Topic: Statistical analysis of diffusion tensors in diffusion-weighted magnetic resonance image data

Date: 12/14/06 (Joint meeting with the working group on Geometric Methods)
Speaker: Makram Talih
Topics: 1. Loss functions on the space of covariance/precision matrices.
2. Matrix-valued martingales/random walks

Date: 12/14/06 (Joint meeting with the working group on Geometric Methods)
Speaker: Donald Richards
Topics: Diffusion tensor imaging, and deconvolution density estimation on spaces of positive definite symmetric matrices

5.7 *Graphical Models/Bayesian Methods*

Group Leader: Helene Massam
Webmaster: Bala Rajaratnam

5.7.1 Identification of issues of interest

From the opening workshop that took place from September 17th to 20th, 2006, it was clear that research in the area of graphical models had common threads with several of the research areas of various speakers at the Opening Workshop. The following most important common threads were identified:

(a) Moment computations for Wishart matrices. H. Massam gave a talk on this topic at the opening workshop. The results are closely linked to the work presented the following day by James Mingo and Raj Rao.

(b) Estimation of large random matrices. Dealing with large random matrices is why statistical graphical models were created. One of the advantages of graphical models is the reduced number of parameters. This is clearly linked to the concept of regularization, within the broader area of high-dimensional inference, which was at the heart of at least two talks, by Bickel and Levina and by La_erty and Wasserman.

(c) The study of various extensions of the Wishart distribution. Extensions and modifications of the Wishart distribution arise naturally in the context of graphical models. The fascinating talk by Ioanna Dimitriu gave us a glimpse in the various β-ensembles of random matrices and the behaviour of the corresponding eigenvalues. Further studies of the Wishart extensions in statistics will certainly benefit from results already obtained by researchers such as Dimitriu.

(d) Last but certainly not least, model selection which was the common underlying thread to many of the "applied" talks.

Parameter estimation and model selection problems are at the heart of much of the work in Bayesian statistics and the combination of graphical models and Bayesian statistics was a natural one. The topics mentioned above were therefore identified at the end of the Opening Workshop as some of the possible areas of research for our GMBM group.

5.7.2 The participants

Twenty six people signed up to be part of our group and were present during the weekly meetings on various occasions. The regular participants were Eitan Greenshtein, Jinchi Lv, Hélène Massam, Debashis Paul, Bala Rajaratnam, Makram Talih and Zheng Lei. Jim Berger and Carlos Carvalho also joined us for many meetings.

5.7.3 Activities

Activities went on at several levels:

1. *Weekly two-hour GMBM meetings* where papers pre-assigned by the group leader were read and commented. The reading was usually directed by the group leader at the beginning but always turned into a deep and lively discussion of the topic considered. As listed on the SAMSI website, the papers read in the first few weeks were
    - Bickel and Levina (2006) Regularized estimation of large covariance matrices
    - Meinhausen and Buhlmann, (2006) High dimensional graphs and variable selection with the LASSO
    - Dobra and West (2004) Bayesian Covariance Selection
    - Jones et al (2005), Experiments in Stochastic Computation for High-Dimensional Graphical Models
    - Wainwright and Jordan (2003)

    The last three weekly meetings were concentrated around the presentation of the paper by Berger and Sun (2006) by Dongchu Sun (two sessions) and a presentation and discussion on reference priors by Jim Berger.

2. *Two workshops*

    - Bayesian Focus Week (oct. 30-Nov.3), organised by Hélène Massam, a week-long workshop with twenty speakers from the US, Canada and Europe

    - Large Graphical Models and Random Matrices (Nov. 9-11, 2006), organized by Nanny Wermuth, a two- and half-day long workshop with seventeen speakers from the US, Canada and Europe.

    Of course, a lot of discussions took place during these two workshops and several close scientific contacts established. The organizers from both workshops received invaluable assistance from B. Rajaratnam.

3. *Research and results*

    At this moment, it is difficult to fully identify all the scholarly work that is or will be done as a consequence of our work at SAMSI during the Fall of 2006. However, we can already say the following three papers are being written or have been written:

    - Bala Rajaratnam, Hélène Massam and Carlos Carvalho are writing a paper on covariance estimation which they intend to submit to the Annals of Statistics for the special SAMSI issue that has been planned.

- Hélène Massam, Debashis Paul and Bala Rajaratnam are writing a paper on model search for discrete log-linear models which should be completed by the summer of 2006.

- Eitan Greenshtein, Junyong Park and Ya'acov Ritov have completed the writing of a paper entitled "Estimating the mean of high valued observations in high dimensions".

It should be noted that the first two papers are the direct product of the cooperation facilitated by the program on Large Random Matrices. The ideas were discussed shortly before the fall but the work was done entirely during this 2006-2007 academic year.

4. *Training of a postdoctoral fellow*

   Bala Rajaratnam is a postdoctoral fellow at SAMSI for the academic year 2006/2007. He is co-sponsored by Iain Johnstone and H´el`ene Massam. During the fall of 2006, he was introduced to the research area of graphical models and Bayesian inference. He now has a good basic working knowledge of this topic and he is forging ahead at a fast pace. I expect him to be an independent member of this research community within a few months.

5. *Links established between different research areas*

   The work described above under "Weekly readings" and "Research and Results" is mostly concerned by topics (b) and (d) described above in "Identification of issues of interest". We have already discussed problems in (a) and (c) that we deem interesting. These and other relevant topics were further discussed during the AIM workshop which took place from April 11 to April 14, 2007.

6 Spring Semester Concentration on Geometry

The continuation began with a four day workshop. Both algorithms and the fundamental mathematical objects computed by the algorithms were stressed. One objective of the workshop was to provide computer scientists, statisticians, probabilists, geometers, and topologists and opportunity to interact. The topics that were covered in this workshop were Geometry and Sparsity, Geometry and Topology, Machine Learning, and Random Matrices and Covariances.

After this initial workshop an intensive working group and class formed focusing on theoretical and applied aspects of toplogical statistics. The main effort in this program has been two fold and the mathematical object focused upon has been persistent homology and persistence diagrams. The first topic the group has addressed is a formal notion of consistency for persistence diagrams. A corollary of this is a proof of convergence of homology inference from a point cloud. This approach improves upon results of Niyogi, Smale, and Weinberger. The more relevant part of the work has been focusing on ringing statistical notions of uncertainty to persistence. Two approaches have been proposed and

are currently being implemented: one based upon bootstrap estimates, the other based on Bayesian density estimation procedures. As of this report, this is an ongoing activity that will continue through the Spring Semester, 2007.

7. Outcomes

The product of the working group activities will primarily be represented by publications, student projects and the initiation of ongoing collaborations.

A key impetus for the program was the idea that a variety of problems in high dimensional statistics would benefit by the infusion of results and techniques from an area of mathematics (random matrix theory) that has long focused on many-variable approximations (the "thermodynamic limit"). Conversely, a goal for the program has been to stimulate new research topics and directions for applied mathematicians.

Peter Bickel is the guest editor in charge of a special issue of the *Annals of Statistics* devoted to articles written by participants in the program and in the workshops. So far, nine articles have been accepted for publication out of fifteen submitted.

Numerous collaborations were initiated during the program. Rajaratnam and Nadler on eigenvalues estimation, Rajaratnam and Levina on L1 Lasso penalization for covariance selection models. Paul and with Raj Rao: on the use of random matrices in signal processing. Massam and Wermuth on factor analysis and graphical models, Massam and Schartzmann on the analysis of data such that data points are positive definite matrices. During the conference, there arose the potential for collaboration with numerous attendees, including: work of Richards with Schwartzman on Central Limit Theorems on heat equations on the space of positive definite matrices, Massam with Nadler on the topic of singular perturbations, and Massam with Mingo on applications of free probability to the non-central Wishart distributions. The AIM workshop also provided the opportunity for discussions between Massam and Mingo on Wishart moment computations, a topic that both have tackled but from different points of view. Mingo is going to spend the Fall of 2007 at the Fields Institute in Toronto and they will carry on teaching each other their respective techniques and points of view.