Mathematisches Forschungsinstitut Oberwolfach

Report No. 41/2014

DOI: 10.4171/OWR/2014/41

New Horizons in Statistical Decision Theory

Organised by
Richard Gill, Leiden
Madalin Guta, Nottingham
Michael Nussbaum, Ithaca

7 September – 13 September 2014

Abstract. The classical metric theory of statistical models (experiments) has recently been extended towards an asymptotic equivalence paradigm, allowing to classify and relate problems which are essentially infinite dimensional and ill-posed. Modern statistical concepts like these are also being integrated into the emerging field of quantum statistics, which is developing on the background of technological breakthroughs in quantum engineering. The workshop brought together leading experts in these areas, with the goal of establishing a common language, and fostering collaborations between mathematical statisticians, theoretical physicists and experimentalists.

Mathematics Subject Classification (2010): 62G20, 81P45.

Introduction by the Organisers

The workshop *New Horizons in Statistical Decision Theory* was the first significant meeting bringing together researchers from mathematical statistics and quantum information theory, under the broad umbrella of statistical decision theory. The aim of the workshop was twofold. The first goal was to review recent progress in these areas, e.g. in non-parametric regression, confidence intervals, quantum local asymptotic normality and quantum compressed sensing tomography. The second, and perhaps more important goal, was to establish a communication platform and facilitate the exchange of methodology and techniques between the two fields.

Recent progress in quantum information technologies has brought the statistical analysis of quantum measurements data to the forefront of experimental and theoretical efforts. The increasing complexity of quantum devices requires a new

range of statistical methods to deal with large dimensional models, model selection, measurement design, and reliable confidence intervals. In the same time, many key statistical concepts from statistical decision theory have been extended to quantum statistics, bringing the two subjects closer together, and making the workshop a very timely event.

In recognition of his pioneering work at the interface of quantum theory, information theory and statistics, the workshop was opened with a presentation by Alexander Holevo on the recently solved quantum Gaussian optimizers conjecture. The program contained a mixture of alternating statistics and quantum information presentations. To increase the accessibility, the speakers observed the "15 minutes rule" of beginning with a broad overview of the subject. Additionally, a lively dictionary session was organised on Tuesday, and several open problems were debated in another session on Thursday. PhD students had the opportunity to present their results with short presentations in a special evening session.

As organisers we were gratified by the level of engagement of participants on both sides, lively discussions and emerging collaborations. The excellent atmosphere was facilitated by the working environment at the MFO to which we would like to express our deep gratitude.

Richard Gill, Madalin Guta and Michael Nussbaum

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1049268, "US Junior Oberwolfach Fellows".

Workshop: New Horizons in Statistical Decision Theory

Table of Contents

Abstracts

Convergence rates of spectral methods for statistical inverse learning problems

Gilles Blanchard
(joint work with Nicole Mücke)

Consider an inverse problem of the form $g = Af$, where $A : \mathcal{H}_1 \to \mathcal{H}_2$ is a known operator between Hilbert spaces of real-valued functions on a space $\mathcal{X}$, and assume that we observe g at some randomly drawn points $X_1, \ldots, X_n$ which are i.i.d. according to some distribution P_X, and where additionally each observation is subject to a random independent noise, i.e.

$$Y_i = (Af)(X_i) + \varepsilon_i, i = 1, \ldots, n.$$

The goal is to recover the function f. Here it is assumed that for each point x the evaluation mapping $f \mapsto Af(x)$ is continuous. This setting as well as its relation to random nonparametric regression and statistical learning with reproducing kernels can be traced back to the works of G. Wahba and has been studied more recently in a series of works by Caponnetto, De Vito, Rosasco, Bauer and Pereverzev. In particular, it can be shown that this setting is geometrically equivalent to the random nonparametric regression model $Y_i = h(X_i) + \varepsilon_i, i = 1, \ldots, n$, wherein it is assumed that h belongs to a certain reproducing kernel Hilbert space $\mathcal{H}_K$ over $\mathcal{X}$, and the goal is to recover the function h with the estimation error being measured in $\mathcal{H}_K$-norm (as opposed to the $L^2(P_X)$ norm for the standard least squares regression setting). In this talk we consider the estimation of f from the observations $(X_i, Y_i)_{1 \leq i \leq n}$ by so-called spectral methods. The results we present concern convergence rates of such methods under Hölder source conditions with parameter r, and polynomial decay condition (with exponent $1/s$) of eigenvalues, both with respect to an appropriate integral operator defined from A and the marginal distribution P_X. These results extend and complete previously known ones, in particular the rate $O(n^{-\frac{r}{2r+s+1}})$ for the convergence in $\mathcal{H}_1$-norm (or equivalently in $\mathcal{H}_K$-norm for the equivalent random design regression model) as well as a corresponding minimax lower bound, both of which had not been established for this setting.

References

[1] A. Caponnetto, *Optimal Rates for Regularization Operators in Learning Theory*, Technical report MIT-CSAIL-TR-2006-062, Massachusetts Institute of Technology (2006).

[2] A. Caponnetto, C. Carmeli, E. De Vito, L. Rosasco, A. Toigo, *Discretization Error Analysis for Tikhonov Regularization*, Analysis and Applications 4 (1) (2006).

[3] F. Bauer, S. Pereverzev, L. Rosasco, *On Regularization Algorithms in Learning Theory*, J. Complexity 23(1): 52-72 (2007)

Gate-set tomography: calibration-free full characterization of quantum devices using error-amplifying circuits

Robin Blume-Kohout
(joint work with John K. Gamble, Peter Maunz, Erik Nielsen, Kenneth Rudinger)

Quantum Information & Tomography: Beginning around 1980, it became apparent that quantum mechanics provides a new and distinct set of rules for information processing. Information is carried by physical systems (e.g., bits), and sufficiently small and/or isolated physical systems obey the rules of quantum theory, *not* those of classical logic and probability theory. (For example, the law of the excluded middle – exactly one of A and not-A is true – simply does not hold for quantum systems.) These rules lead to novel, and sometimes useful, information processing possibilities. Achieving these capabilities relies on the development of *qubits*: physical systems that can be prepared in either of two distinguishable states (labeled $|0\rangle$ and $|1\rangle$) *and* in arbitrary quantum superpositions of them. A useful qubit must be precisely controllable via *quantum logic gates*. These are logical transformations (akin to the classical NOT or NAND gates), usually implemented by applying electronic or optical control pulses to the qubit. The precision required to process quantum information demands precise statistical characterization of these logic gates, a.k.a. *quantum tomography*. This talk is about a new method for tomography of quantum logic gates.

Quantum tomography is deeply similar to estimation of: (i) probability distributions over finite sample spaces, e.g. $\vec{p} = [p_1 \dots p_d]$; and (ii) stochastic matrices that act on probability vectors. The key difference is that quantum systems do not have unique sample spaces. Instead, they can be observed in a continuum of different ways, and each possible observation corresponds to a distinct sample space that is not a coarse-graining of any other. Fortunately, there exist linear relationships between them, which ensure that quantum states for a system with d distinguishable states can be represented as $d \times d$ positive semidefinite *density matrices* (ρ). Similarly, quantum logic gates can be represented by $d^2 \times d^2$ *process matrices* (G), which are linear maps on density matrices.

The standard methods for tomography (i.e., estimation of the process matrix for one or more quantum logic gates) are applications of frame theory. The gate is represented by a $d^2 \times d^2$ matrix G that acts on quantum states, so the tomographer:

(1) devises ways to prepare each of d^2 linearly independent states ρ_i,
(2) repeatedly prepares ρ_i, and applies the logic gate to it,
(3) measures the resulting states $G[\rho_i]$ in enough different ways (one for each sample, since measurement "collapses" quantum states) to obtain information about every parameter of $G[\rho_i]$,
(4) performs statistical analysis on the resulting data to (i) estimate each $G[\rho_i]$, and then (ii) estimate the entire matrix G from them.

This procedure works, but is bedeviled by many challenges. Two of the most significant are: (1) it relies on *known* input states and measurements, and is only as reliable and accurate as their calibration; (2) an enormous amount of data is

required to achieve the desired accuracy of 10^{-4} or better in every parameter, because errors scale as $1/\sqrt{N}$.

Gate-set Tomography: Recent experimental work [1] brought the calibration problem into stark relief. In response, various researchers (notably at IBM [2]) devised ad-hoc methods for self-calibrating tomography. We developed a rigorous method called *gate set tomography* (GST) [3]. It completely solves the calibration problem, by treating the quantum logic device (e.g. qubit) as a black box, accessible only through buttons and indicator lights whose meaning and function are strictly unknown to begin. Remarkably, it also provides a tremendous advantage in *efficiency*; simulations show that GST can estimate every parameter of a set of logic gates to within 10^{-6} using only 5×10^6 samples.

GST models an experimental quantum logic device as a black box, equipped with buttons (to control it) and indicator lights (one of which lights up after the "measure" button is pressed). Three kinds of buttons exist: (1) one to initialize in an unknown state ρ; (2) several to apply unknown logic gates $G_1 \dots G_K$; and (3) one to perform an unknown two-outcome measurement $\mathcal{M} = \{E, \mathbb{1} - E\}$. ρ and E are $d \times d$ positive semidefinite matrices (with $E \leq \mathbb{1}$ and $\mathrm{Tr}(\rho) = 1$), and the G_k are completely positive trace preserving (CPTP) linear maps acting on ρ. Together, these parameters constitute a *gateset*, a complete description of the device's behavior, which is parameterized using the Hilbert-Schmidt vector space of Hermitian matrices (in which states are row vectors, measurement effects are column vectors, and gates/operations are matrices) as:

$$\{|\rho\rangle\rangle, \langle\langle E|, \{G_k\}\}.$$

The goal of GST is to estimate these parameters as accurately as possible. Actually, this is not quite possible; there is an unobservable *gauge* in the gateset. In quantum theory, Born's Rule gives the outcome probabilities of each possible experiment by an expression of the form

$$\Pr(E|\rho, G_1 \dots G_L) = \langle\langle E| G_L \dots G_1 |\rho\rangle\rangle, \tag{1}$$

which is unchanged under gauge transformations of the form

$$\langle\langle E| \to \langle\langle E| \mathbf{T}^{-1}, \ |\rho\rangle\rangle \to \mathbf{T}|\rho\rangle\rangle, \ G_k \to \mathbf{T} G_k \mathbf{T}^{-1}. \tag{2}$$

Thus, the gateset can only be estimated up to this gauge degree of freedom, which GST generally fixes by choosing the gauge in which the gateset is closest to the experimentalist's intended gates.

Results & Specifics: We have developed a comprehensive suite of algorithms and computer code for GST on single-qubit systems. (Extensions to larger systems, with Hilbert space dimension $d > 2$, are in progress and expected to be straightforward). At its heart is an algorithm called *linear gate set tomography* (LGST). Previous methods (e.g. Ref. [2]) used maximum likelihood to analyze data from arbitrary experiments (each of which is described by a *gate sequence* $G_1 \dots G_L$). This method is risky because the likelihood function is complicated,

non-convex, and multimodal. LGST avoids these difficulties by specifying a particular set of short-sequence experiments whose results can be transformed into a consistent estimator using only simple linear algebra.

LGST has near-perfect reliability, but is statistically naive and not very accurate. Long gate sequences, in which a short sequence of gates or *germ* is repeated many times, amplify small deviations in the gates, and make them easy to detect and estimate. These data cannot be analyzed with linear methods. However, the linear inversion used in LGST equates to unweighted least-squares fitting, and properly weighted least-squares (minimizing χ^2) approximates maximum likelihood estimation in the Gaussian limit. We synthesized these ideas into *least squares GST* (LSGST), which specifies particular *long* gate sequences, and analyzes the data using iterative weighted least-squares refinement of the initial LGST estimate.

Using these methods, we have analyzed a variety of *simulated* data, to validate and calibrate our methods. We have also analyzed *experimental* data from three different laboratories. Our simulations showed that LSGST reliably yields an estimate whose error (RMS per-matrix-element deviation from the truth) scales precisely as $1/L$, where L is the maximum length of gate sequences used. For $L = 8192$, using a total of about 5×10^6 "clicks" (individual samples) we observed errors of 9×10^{-7}.

We analyzed experimental data with sequences of length up to $L = 512$. GST produced estimates with far higher precision than any existing technique, thanks to the $1/L$ error scaling, and an algorithm to design sets of germs that, when repeated L times, amplify every gauge-invariant parameter proportional to L. We estimated these parameters to within 10^{-4}, and discovered that the dominant source of noise in all three experimental qubits was *non-Markovian* – i.e., violation of the [Markovian] gate-set model. We used a detailed χ^2 analysis (which assigned a χ^2 badness-of-fit number to each one of the roughly 2000 different experiments performed) to obtained detailed diagnoses of the non-Markovian noise, and this information enabled some experimentalists to improve their qubits.

Acknowledgments: Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

[1] M. D. Schulman *et al*, *Demonstration of entanglement of electrostatically coupled singlet-triplet qubits*, Science **336** (2012), 202–205.

[2] S. T. Merkel *et al*, *Self-consistent quantum process tomography*, Phys. Rev. A **87** (2013), 062119.

[3] R. Blume-Kohout *et al*, *Robust, self-consistent, closed-form tomography of quantum logic gates on a trapped ion qubit*, `arxiv.org:1310.4492` (2013).

Sparse priors and Bayesian linear regression

Ismaël Castillo
(joint work with Johannes Schmidt-Hieber, Aad van der Vaart)

Consider estimation of a parameter $\beta \in \mathbb{R}^p$ in the linear regression model

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

where X is a given, deterministic $(n \times p)$ matrix, and ϵ is an n-variate standard normal vector. We are interested in the *sparse* setup, where possibly $n \ll p$, and 'many' or 'most' of the coefficients β_i of the parameter vector are zero, or close to zero. We study a Bayesian approach based on a prior distribution Π that sets a selection of coefficients β_i a priori to zero: the behaviour of the posterior distribution $\Pi[\cdot \,|\, Y]$ is investigated under the 'frequentist' assumption that the data Y has in reality been generated according to a given sparse parameter β^0.

Specifically, we consider a prior Π on β that first selects a *dimension* s from a prior π_p on the set $\{0, \ldots, p\}$, next a random subset $S \subset \{1, 2, \ldots, p\}$ of cardinality $|S| = s$, and finally a set of nonzero values $\beta_S := \{\beta_i : i \in S\}$ from a prior density g_S on $\mathbb{R}^S$. Formally, the prior on (S, β) can be expressed as

$$(S, \beta) \mapsto \pi_p(|S|) \frac{1}{\binom{p}{|S|}} g_S(\beta_S) \delta_0(\beta_{S^c}), \tag{2}$$

where $\delta_0(\beta_{S^c})$ refers to the coordinates $\beta_{S^c} := (\beta_i : i \in S^c)$ being zero. We focus on the situation where g_S is a product $\otimes g$ of densities over the coordinates in S, for g the Laplace density on $\mathbb{R}$ with parameter λ that can be chosen in the range

$$\frac{\|X\|}{p} \leq \lambda \leq 4\|X\| \sqrt{\log p}, \tag{3}$$

where $\|X\| = \max_{i=1,\ldots,p} (X^t X)_{i,i}^{1/2}$. This is a natural continuation of [2], that considered the special case where X is the identity matrix and $p = n$. The general model (1) is different in that it must take account of the noninvertibility of X and its interplay with the sparsity assumption, and does not allow a factorization of the model along the coordinate axes.

To overcome the nonidentifiability of the full parameter vector β in the over-specified model (1) we borrow from the work on sparse regression within the non-Bayesian framework, such as [5], [3], [4]. Good performance of the posterior distribution is shown under *compatibility* and *smallest sparse eigenvalue* conditions

For a subset S of $\{1, \ldots, p\}$, we define its compatibility number $\phi(S)$ by

$$\phi(S) = \inf_{\|\beta_{S^c}\|_1 \leq 7\|\beta\|_1} \frac{\|X\beta\|_2 \sqrt{|S|}}{\|X\| \|\beta_S\|_1}.$$

We say that compatibility in s_n-sparse vectors holds if

$$\inf_{\beta:\, \|\beta\|_0 \leq 5s_n} \frac{\|X\beta\|_2 \sqrt{|S_\beta|}}{\|X\| \|\beta\|_1} \gg 0.$$

Recovery. Our main result for estimation of β is as follows. Set $\phi(\beta_0) := \phi(S_{\beta_0})$.

Theorem 1. *Under compatibility of s_n-sparse vectors, for every $c > 0$,*

$$\sup_{\|\beta_0\|_0 \le s_n, \phi(\beta_0) \ge c} E_{\beta_0} \Pi[\, \beta,\ \|X(\beta - \beta_0)\|_2 \gtrsim \sqrt{s_n \log p} \,|\, Y] \to 0$$

$$\sup_{\|\beta_0\|_0 \le s_n, \phi(\beta_0) \ge c} E_{\beta_0} \Pi[\, \beta,\ \|\beta - \beta_0\|_1 \gtrsim s_n \sqrt{\log p}/\|X\| \,|\, Y] \to 0$$

In words, the posterior distribution achieves the minimax rate for estimation of β in terms of the $\|\cdot\|_1$-norm over sparse classes. Similar results can be obtained for $\|\cdot\|_2$ and $\|\cdot\|_\infty$ norms, under slightly stronger conditions on X, in line with results from the frequentist literature.

Model selection. If all nonzero coefficients of β_0 are appropriately large (depending on what is assumed on the matrix X), then we can show that the posterior asymptotically recovers the true model in that $\Pi[S = S_{\beta_0} \,|\, Y] \to 1$.

Prediction. If estimation of $X\beta$ rather than β is the main goal, then no conditions on X (e.g. compatibility) should be necessary. For a Bayesian-flavoured method (a pseudo-posterior mean estimator) it was shown in [6] that it was indeed possible to achieve (nearly-)minimax rates for prediction with arbitrary X. We find a similar result for the full posterior distribution for priors of the type (2).

Limiting shape of the posterior. Our second main result considers the asymptotic shape of the posterior distribution for 'small λ' in (3), namely for

$$\frac{\lambda}{\|X\|} s_n \sqrt{\log p} \to 0, \tag{4}$$

which can be seen as asking for a 'flat' prior on nonzero-coordinates. In the next statement, $\|\cdot - \cdot\|$ is the total variation distance between measures.

Theorem 2. *Under compatibility for s_n-sparse vectors, for λ as in* (4)*,*

$$E_{\beta_0} \| \, \Pi[\cdot \,|\, Y] - \sum_S \hat{w}_S N(\hat{\beta}_{(S)}, \Gamma_S^{-1}) \otimes \delta_{S^c} \, \| \to 0$$

with $\hat{\beta}_{(S)}$ least square estimate in model S, with covariance Γ_S^{-1} and

$$\hat{w}_S \propto \frac{\pi_p(s)}{\binom{p}{s}} \Big(\frac{\lambda\sqrt{2\pi}}{2} \Big)^s |\Gamma_S|^{-1/2} e^{\frac{1}{2}\|X_S \hat{\beta}_{(S)}\|_2^2} \, 1_{|S| \le 4s_n,\, \|\beta_{0,S^c}\|_1 \lesssim s_n \sqrt{\log p}/\|X\|}.$$

The limiting distribution in Theorem 2 is thus a (random) mixture of normal distributions of least-squares estimators over specific submodels described in the indicator function appearing in the expression of the weights $\hat{w}_S$. Under conditions ensuring exact model selection as discussed above, this mixture degenerates into a single normal law, the asymptotic distribution of $\hat{\beta}_{(S_0)}$,

$$E_{\beta_0} \| \Pi[\cdot \,|\, Y] - N(\hat{\beta}_{(S_0)}, \Gamma_{S_0}^{-1}) \| \to 0.$$

A consequence of this convergence is that, under signal strength conditions guaranteeing the above result, credible sets for a given coefficient β_i are asymptotic confidence sets.

A word on simulations. We note that, although computationally more involved than methods such as the LASSO, simulation from posterior distributions –or aspects of it– such as the one considered here has recently attracted quite a lot of attention and promising results have so far been obtained for moderate p (up to a few thousands). Recent works on the subject include [8], [9], [7].

References

[1] I. Castillo, J. Schmidt-Hieber and A. van der Vaart, *Bayesian linear regression with sparse priors*, Preprint arXiv:1403.0735, March 2014.

[2] I. Castillo and A. van der Vaart, *Needles and straw in a haystack: posterior concentration for possibly sparse sequences*, *Ann. Statist.*, 40(4):2069–2101, 2012.

[3] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[4] P. Bühlmann and S. van de Geer, *Statistics for High-dimensional Data*, Springer, Berlin (2011)

[5] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[6] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 97–111. Springer, Berlin, 2007.

[7] R. Martin, R. Mess, and S. G. Walker. Empirical Bayes posterior concentration in sparse high-dimensional linear models. *ArXiv e-prints*, June 2014.

[8] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011.

[9] A. Schreck, G. Fort, S. Le Corff, and E. Moulines. A shrinkage-thresholding Metropolis adjusted Langevin algorithm for Bayesian variable selection. *ArXiv e-prints*, Dec. 2013.

Bayes estimator of Bhattacharyya loss via the quantum route

Christopher Ferrie

(joint work with Robin Blume-Kohout)

In statistical decision theory, the Bayes estimators are point estimators which have average risk optimality properties [1]. The risk is defined through the specification of a loss function and Bayes estimator is sensitive to this choice. Many common loss functions have a Bayes estimator which is the mean of posterior distribution. Generally, Bayes estimators are useful not only to understand Bayeisan optimality, but they also provide lower bounds the frequentist concept of minimax risk, such that the greatest lower bound coincides with the frequentist risk. This duality leads to, for example, numerically efficient algorithms for finding minimax estimator by iterating over Bayesian priors, rather than searching the computationally intractable space of estimators [2].

Here we will consider a loss function defined through the Bhattacharyya coefficient, a distinguishability measure on probability distributions. It is often used to

define distances between probability distributions for applications in, for example, machine learning (e.g. [3]). In quantum information theory, it is used to define the *fidelity* [4, 5], which is the most commonly used distinguishability measure in both theoretical and experimental studies. It is surprising then that only in a few special cases is the Bayes estimator actually known. On the other hand, we do not solve the fully general quantum mechanical problem here. Our result, however, is sufficiently general to solve its classical analogue.

Suppose we have a die or a finite alphabet or some other finite set whose elements are selected independently and we wish to estimate the probabilities associated to each single trial outcome. If each trial has K potential outcomes, then the probabilities form a vector $\vec{p}$ which belongs to the K-simplex

$$\triangle_K = \left\{ (p_0, \ldots, p_{K-1}) \middle| p_k \geq 0 \ \forall k, \sum_k p_k = 1 \right\}. \tag{1}$$

For us, it is really not that important how the data are generated from $\vec{p}$. Typically, $\vec{p}$ are considered the bare probabilities for the K events, but we could also have noisy observations or correlations which are not fundamentally related to the properties of the die but particular to the way in which the data was generated or recorded. In any case, we will not find it important what the function form of the likelihood function $\Pr(\texttt{data}|\vec{p})$ is. What is important in the Bayesian decision theoretic framework is the posterior $\Pr(\vec{p}|\texttt{data})$, as we will see next.

From the data, we wish to produce an estimate of $\vec{p}$, call it $\hat{\vec{p}}$. We want the estimate to be "good". This is formalized through a loss function $L(\vec{p}, \hat{\vec{p}})$, which gives us a numerical value for "how bad" it is to give an estimate $\hat{\vec{p}}$ when the truth is $\vec{p}$. It makes sense then to suppose that $L(\vec{p}, \vec{p}) \leq L(\vec{p}, \hat{\vec{p}})$, such that our task becomes minimizing the loss function.

There is still a problem in that both the data and $\vec{p}$ are random variables. The Bayesian solution is to average the loss over the posterior resulting in the *Bayes risk* of $\hat{\vec{p}}$:

$$r(\hat{\vec{p}}) = \mathbb{E}_{\vec{p},\texttt{data}}[L(\vec{p}, \hat{\vec{p}}(\texttt{data})]. \tag{2}$$

An estimator which minimizes the Bayes risk is called a *Bayes estimator* and we will label it $\hat{\vec{p}}_{\mathrm{B}}$. This estimator depends highly on the loss function one uses. Consider, for example, loss functions of the form

$$L(\vec{p}, \hat{\vec{p}}) = (\vec{p} - \hat{\vec{p}})^{\mathrm{T}} \Sigma (\vec{p} - \hat{\vec{p}}), \tag{3}$$

where $\Sigma > 0$ is a positive definite matrix. For this *quadratic* loss function the Bayes estimator is the mean of the posterior distribution [6]

$$\hat{\vec{p}}_{\mathrm{B}}(\texttt{data}) = \mathbb{E}_{\vec{p}|\texttt{data}}[\vec{p}]. \tag{4}$$

As we noted above, the data themselves and the distribution from which they were generated are not important once the posterior distribution has been calculated and we will drop this conditional information from now on. Another important loss

function is the *relative entropy* (also known as the Kullback-Leibler divergence):

$$L(\vec{p}, \hat{\vec{p}}) = \sum_k p_k \log \frac{p_k}{\hat{p}_k}. \tag{5}$$

Again, the Bayes estimator for this loss function is the mean of the posterior distribution [7].

The loss function we consider here is related to the Bhattacharyya coefficient, B,

$$B(\vec{p}, \hat{\vec{p}}) = \sum_k \sqrt{p_k \hat{p}_k}, \tag{6}$$

as follows

$$L(\vec{p}, \hat{\vec{p}}) = 1 - B(\vec{p}, \hat{\vec{p}})^2. \tag{7}$$

The particular form we choose (one minus the square) is inherited from quantum information theory. And, using introductory quantum information theoretic techniques [8], we show that the Bayes estimator for this loss function is

$$\hat{\vec{p}} = (a_0^2, a_1^2, \ldots, a_{K-1}^2), \tag{8}$$

where the vector $\vec{a} = (a_0, a_1, \ldots, a_{K-1})$ is the eigenvector associated with the maximal eigenvalue of the matrix with entries $\mathbb{E}_{\vec{p}}[\sqrt{p_j p_k}]$.

References

[1] J. Berger, *Statistical Decision Theory and Bayesian Analysis*, Springer (1985).

[2] P. J. Kempthorne, *Numerical specification of discrete least favorable prior distributions*, SIAM Journal on Scientific and Statistical Computing **8**, 171 (1987).

[3] A. Djouadi, O. Snorrason and F. D. Garber, *The quality of training sample estimates of the Bhattacharyya coefficient*, IEEE Transactions on Pattern Analysis and Machine Intelligence **12**, 92 (1990).

[4] W. K. Wootters, *Statistical distance and Hilbert space*, Physical Review D **23**, 357 (1981).

[5] R. Jozsa, *Fidelity for mixed quantum states*, Journal of Modern Optics **41**, 2315 (1994).

[6] E. L. Lehmann and G. Casella, *Theory of point estimation*, Springer (1998).

[7] J. Aitchison, *Goodness of prediction fit*, Biometrika **62**, 547 (1975).

[8] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*. Cambridge University Press (2010).

Weak local asymptotic normality in the quantum domain

Akio Fujiwara

(joint work with Koichi Yamagata, Richard D. Gill)

Suppose that one has n copies of a quantum system each in the same state depending on an unknown parameter θ, and one wishes to estimate θ by making some measurement on the n systems together. Given the measurement, we have a classical parametric statistical model, though not necessarily an i.i.d. model, since we are allowed to bring the n systems together before measuring the resulting joint system as one quantum object. A question naturally arises: what is the best we can do as $n \to \infty$? The objective of this work is to study this question by

extending the theory of local asymptotic normality (LAN) to quantum statistical models. For details, consult [9]. (See also [1, 2, 3, 4, 5, 8] for related works.)

Given a $d \times d$ real skew-symmetric matrix $S = [S_{ij}]$, let $\mathrm{CCR}(S)$ be the algebra generated by a set of observables $X := (X_1, \ldots, X_d)$ satisfying the canonical commutation relation (CCR):

$$\frac{\sqrt{-1}}{2}[X_i, X_j] = S_{ij} I \qquad (1 \le i, j \le d).$$

A state ϕ on the algebra $\mathrm{CCR}(S)$ is characterized by the *characteristic function*

$$\mathcal{F}_\xi\{\phi\} := \phi(e^{\sqrt{-1}\xi^i X_i}),$$

where $\xi = (\xi^i)_{i=1}^d \in \mathbb{R}^d$. A state ϕ on $\mathrm{CCR}(S)$ is called a *quantum Gaussian state* [6], denoted by $\phi \sim N(h, J)$, if the characteristic function takes the form

$$\mathcal{F}_\xi\{\phi\} = e^{\sqrt{-1}\xi^i h_i - \frac{1}{2}\xi^i \xi^j V_{ij}},$$

where $h = (h_i)_{i=1}^d \in \mathbb{R}^d$ and $V = (V_{ij})$ is a real symmetric matrix such that the Hermitian matrix $J := V + \sqrt{-1}S$ is positive semidefinite. When the canonical observables X need to be specified, we also use the notation $(X, \phi) \sim N(h, J)$.

Suppose we are given, for each $n \in \mathbb{N}$, a density operator $\rho^{(n)}$ and a list of observables $X^{(n)} := (X_1^{(n)}, \ldots, X_d^{(n)})$ on a Hilbert space $\mathcal{H}^{(n)}$. We say the sequence $\left(X^{(n)}, \rho^{(n)}\right)$ *converges in law to a quantum Gaussian state* $(X, \phi) \sim N(h, J)$, denoted as $(X^{(n)}, \rho^{(n)}) \underset{q}{\rightsquigarrow} N(h, J)$, if the quasi-characteristic function [7] of $(X^{(n)}, \rho^{(n)})$ converges to that of $N(h, J)$, that is,

$$\lim_{n\to\infty} \mathrm{Tr}\, \rho^{(n)} \left(\prod_{t=1}^r e^{\sqrt{-1}\xi_t^i X_i^{(n)}} \right) = \phi \left(\prod_{t=1}^r e^{\sqrt{-1}\xi_t^i X_i} \right),$$

for any finite subset $\{\xi_t\}_{t=1}^r$ of $\mathbb{C}^d$.

We say a pair of density operators ρ and σ on a Hilbert space $\mathcal{H}$ are *mutually absolutely continuous,* $\rho \sim \sigma$ in symbols, if there exists a selfadjoint operator $\mathcal{L}$ that satisfies

$$\sigma = e^{\frac{1}{2}\mathcal{L}} \rho\, e^{\frac{1}{2}\mathcal{L}}.$$

We shall call such a selfadjoint operator $\mathcal{L}$ a *quantum log-likelihood ratio.* When the reference states ρ and σ need to be specified, $\mathcal{L}$ shall be denoted by $\mathcal{L}(\sigma|\rho)$.

A sequence of quantum statistical models $\{\rho_\theta^{(n)} \,;\, \theta \in \Theta \subset \mathbb{R}^d\}$, each defined on a finite dimensional Hilbert space $\mathcal{H}^{(n)}$, is called *quantum locally asymptotically normal* (QLAN) at $\theta_0 \in \Theta$ if $\rho_\theta^{(n)} \sim \rho_{\theta_0}^{(n)}$ for all $\theta \in \Theta$ and $n \in \mathbb{N}$, and the quantum log-likelihood ratio $\mathcal{L}_h^{(n)} := \mathcal{L}\left(\rho_{\theta_0+h/\sqrt{n}}^{(n)} \middle| \rho_{\theta_0}^{(n)}\right)$ is expanded in $h \in \mathbb{R}^d$ as

$$\mathcal{L}_h^{(n)} = h^i \Delta_i^{(n)} - \frac{1}{2}(J_{ij} h^i h^j) I^{(n)} + o(\Delta^{(n)}, \rho_{\theta_0}^{(n)}).$$

Here $\Delta^{(n)} := (\Delta_1^{(n)}, \ldots, \Delta_d^{(n)})$ is a list of observables on $\mathcal{H}^{(n)}$ satisfying

$$\left(\Delta^{(n)}, \rho_{\theta_0}^{(n)}\right) \underset{q}{\rightsquigarrow} N(0, J),$$

$I^{(n)}$ the identity operator on $\mathcal{H}^{(n)}$, and $o(\Delta^{(n)}, \rho_{\theta_0}^{(n)})$ denotes an infinitesimal residual term in view of the convergence of quasi-characteristic function [9].

To formulate a quantum extension of Le Cam's third lemma, we need a device to handle the infinitesimal residual term in a more elaborate way. Let $X^{(n)} = (X_1^{(n)}, \ldots, X_r^{(n)})$ be a list of observables on $\mathcal{H}^{(n)}$. We say the pair $(\rho_\theta^{(n)}, X^{(n)})$ is *jointly QLAN* at $\theta_0 \in \Theta$ if $\rho_\theta^{(n)} \sim \rho_{\theta_0}^{(n)}$ for all $\theta \in \Theta$ and $n \in \mathbb{N}$, and $\mathcal{L}_h^{(n)}$ is expanded in $h \in \mathbb{R}^d$ as

$$\mathcal{L}_h^{(n)} = h^i \Delta_i^{(n)} - \frac{1}{2}(J_{ij} h^i h^j) I^{(n)} + o\left(\begin{pmatrix} X^{(n)} \\ \Delta^{(n)} \end{pmatrix}, \rho_{\theta_0}^{(n)}\right).$$

Here $\Delta^{(n)} = (\Delta_1^{(n)}, \ldots, \Delta_d^{(n)})$ is a list of observables on $\mathcal{H}^{(n)}$ satisfying

$$\left(\begin{pmatrix} X^{(n)} \\ \Delta^{(n)} \end{pmatrix}, \rho_{\theta_0}^{(n)}\right) \underset{q}{\rightsquigarrow} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^* & J \end{pmatrix}\right),$$

Σ and J are Hermitian positive semidefinite matrices of size $r \times r$ and $d \times d$, and τ is a complex matrix of size $r \times d$. With this assumption, we can prove the following quantum version of Le Cam's third lemma.

Theorem 1. *If $(\rho_\theta^{(n)}, X^{(n)})$ is jointly QLAN at $\theta_0 \in \Theta$, then*

$$\left(X^{(n)}, \rho_{\theta_0 + h/\sqrt{n}}^{(n)}\right) \underset{q}{\rightsquigarrow} N((\operatorname{Re}\tau)\, h, \Sigma)$$

for all $h \in \mathbb{R}^d$.

In applications, we often handle i.i.d. extensions of a given quantum statistical model $\left\{\rho_\theta\,;\ \theta \in \Theta \subset \mathbb{R}^d\right\}$ on a finite dimensional Hilbert space $\mathcal{H}$. In this case we have the following strong result.

Theorem 2. *If $\rho_\theta \sim \rho_{\theta_0}$ for all $\theta \in \Theta$ and $\mathcal{L}(\rho_\theta | \rho_{\theta_0})$ is sufficiently smooth in θ, then $\{\rho_\theta^{\otimes n}\}_\theta$ is QLAN at θ_0, in that $\rho_\theta^{\otimes n} \sim \rho_{\theta_0}^{\otimes n}$, and*

$$\Delta_i^{(n)} := \frac{1}{\sqrt{n}} \sum_{k=1}^{n} I^{\otimes(k-1)} \otimes L_i \otimes I^{\otimes(n-k)}, \qquad (i = 1, \ldots, d),$$

with L_i being the ith symmetric logarithmic derivative (SLD) [6] of ρ_θ at $\theta_0 \in \Theta$, satisfy the requirement for QLAN. Moreover, given arbitrary observables $\{B_i\}_{1 \le i \le r}$ on $\mathcal{H}$ satisfying $\operatorname{Tr}\rho_{\theta_0} B_i = 0$ for $i = 1, \ldots, r$, the pair $(\rho_\theta^{\otimes n}, X^{(n)})$, with

$$X_i^{(n)} := \frac{1}{\sqrt{n}} \sum_{k=1}^{n} I^{\otimes(k-1)} \otimes B_i \otimes I^{\otimes(n-k)}, \qquad (i = 1, \ldots, r),$$

is jointly QLAN at θ_0

An immediate consequence of Theorems 1 and 2 is the following i.i.d. version of the quantum Le Cam third lemma.

Corollary 1. *Given any set of observables* $\{B_i\}_{1\le i\le r}$ *on* $\mathcal{H}$ *satisfying* $\operatorname{Tr}\rho_{\theta_0}B_i = 0$ *for* $i = 1,\dots,r$, *let* $X_i^{(n)}$ *be as in Theorem 2. Then*

$$\left(X^{(n)}, \rho_{\theta_0+h/\sqrt{n}}^{\otimes n}\right) \underset{q}{\rightsquigarrow} N((\operatorname{Re}\tau)\,h, \Sigma)$$

for all $h \in \mathbb{R}^d$, *where* $\Sigma_{ij} = \operatorname{Tr}\rho_{\theta_0}B_jB_i$ *and* $\tau_{ij} = \operatorname{Tr}\rho_{\theta_0}L_jB_i$ *with* L_i *being the* i*th SLD at* θ_0.

It is crucial to observe that the choice of operators $\{B_i\}_{1\le i\le r}$ in Corollary 1 is arbitrary. To put it differently, one can design the limiting quantum Gaussian shift model $N((\operatorname{Re}\tau)\,h, \Sigma)$ at will. This fact plays a key role in proving the asymptotic achievability of the Holevo bound [9].

References

[1] Gill, R. D. and Guţă, M., On Asymptotic Quantum Statistical Inference. *IMS Collections From Probability to Statistics and Back: High-Dimensional Models and Processes* **9** (2012) 105–127.

[2] Guţă, M and Butucea, C., Quantum U-statistics, *J. Math. Phys.* **51** (2010) 102202.

[3] Guţă, M and Jençová, A., Local asymptotic normality in quantum statistics. *Commun. Math. Phys.* **276** (2007) 341–379.

[4] Guţă, M. and Kahn, J., Local asymptotic normality for qubit states. *Phys. Rev. A* **73** (2006) 052108.

[5] Hayashi, M. and Matsumoto, K., Asymptotic performance of optimal state estimation in qubit system. *J. Math. Phys.* **49** (2008) 102101.

[6] Holevo, A. S., *Probabilistic and Statistical Aspects of Quantum Theory*, 2nd English edition (Edizioni della Normale, Pisa, 2011).

[7] Jakšić, V., Pautrat, Y., and Pillet, C.-A., A quantum central limit theorem for sums of independent identically distributed random variables. *J. Math. Phys.* **51** (2010) 015208.

[8] Kahn, J. and Guţă, M., Local asymptotic normality for finite dimensional quantum systems. *Commun. Math. Phys.* **289** (2009) 597–652.

[9] Yamagata, K., Fujiwara, A., and Gill, R. D., Quantum local asymptotic normality based on a new quantum likelihood ratio. *Ann. Statist.* **41** (2013) 2197–2217; Supplementary material (doi: 10.1214/13-AOS1147SUPP).

The fundamental problem of forensic statistics (sparsity, and "less is more")

Richard Gill

(joint work with Dragi Anevski, Stefan Zohren, Maikel Bargpeter, Giulia Cereda)

I report here on joint work with Dragi Anevski and Stefan Zohren, see Anevski, Gill and Zohren (2013), and as yet unpublished work with master student Maikel Bargpeter and with PhD student Giulia Cereda.

The slides of my talk can be found on "slideshare": *http://www.slideshare.net/gill1109/a-walk-in-the-black-forest-during-which-i-explain-the-fundamental-problem-of-forensic-statistics*.

Just recently, Piet Groeneboom has got interested in this subject and I expect him to publish some interesting new results in the near future.

Roughly speaking, the "fundamental problem of forensic statistics" is to give a decent estimate of $1/p_s$, or more meaningfully, of $-\log_{10}(p_s)$, where p_s is the probability of an event which has not yet been observed. A point estimate is not enough, we must also quantify the imprecision of our estimate. The subscript s stands for "species". We are given a sample of animals or other organisms, each animal in the sample belongs to some species, and we are interested in the probability of a particular species (presumably, rare) which did not turn up in our sample. The number of possible species is very large. In forensic science, the "species" (plural) in question would typically be DNA profiles, not of autosomal DNA but, for instance, of Y-chromosome or mitochondrial DNA. Because of the absence of recombination, Y-chromosome profiles (called Y-STR profiles, see `http://yhrd.org`) are distributed with a small number of distinct profiles having quite high probabilities (the most common has probability one in twenty) and very many distinct profiles having rather small probabilities. Imagine now we have a data-base, thought of as a random sample of size a few thousands, from a population of interest; and we have a new case, in which a DNA trace found at the scene of the crime matches the DNA profile of a primary suspect, identified on other grounds. The specific profile in the case does not occur in our data-base so it is rare; the fact of the match between crime-scene and suspect is strong evidence against our suspect. But how strong?

The term "fundamental problem of forensic mathematics" was coined by forensic scientist Charles Brenner who introduced his own solution to this problem, inspired by the famous Good-Turing estimator. The Good-Turing problem is to estimate the probability that when we augment our data-base sample with one new element, it will turn out to belong to a new species (a species not yet observed in the data-base).

More formally, we have a *database* of size n, modelled as a single observation $\mathbf{X} \sim \text{Multinom}(n, \mathbf{p})$ where $\mathbf{p} = (p_s : s \in \mathcal{S})$, $\mathbf{X} = (X_s : s \in \mathcal{S})$, $\mathcal{S}$ is a very large set of *species*, and p_s is the probability of an animal of species s. We are interested in a particular species s such that $X_s = 0$, this is the new species which has turned up in a new crime case. There is a large literature in forensic statistics describing various different approaches to this problem, none of them very satisfying from the mathematical statistical point of view.

In my talk, I described some new approaches to the problem based on two sets of ideas. The first set of ideas is based on the Good-Turing approach to estimating probabilities related to extending the sample from size n to $n + 1$. The famous Good-Turing estimator of the probability that the new item would belong to a different species from all previously observed is just the relative frequency in the data-base of animals of species which are only observed once. The second set of ideas comes from methodology introduced by computer-scientist Alon Orlitsky, in the context of information theory, coding and transmission, in which he proposes to estimate $\mathbf{q} = \text{rsort}(\mathbf{p})$ where "rsort" stands for "reverse sort" or "sort in decreasing order", by *first* reducing the data $\mathbf{X}$ to the statistic $\mathbf{Y} = \text{rsort}(\mathbf{X})$, and *then* applying the maximum likelihood principle. This typically results in a very

different estimate of $\mathbf{q}$ than is obtained by first applying maximum likelihood, then applying the functional. In other words, the maximum likelihood estimator of $\mathbf{q}$ based on the reduced data $\mathbf{Y}$ is typically not equal to $\mathrm{rsort}(\mathbf{X}/n)$.

We call the latter estimator, the *naive estimator* of $\mathbf{q}$. Of course it is not so naive at all to estimate a functional of a parameter by the same functional of the MLE; however, it seems that in the present context we encounter an estimation problem, including the particular functionals of interest, where the "naive" approach evidently gives bad answers. The reason for the bad behaviour is that we are in a typical "sparsity" problem: many parameters, not much data. We need to use extra, sparsity related information ... or come up with some other clever idea.

$\mathbf{Y} = \mathrm{rsort}(X)$ is called the *profile*, *pattern* or *spectrum* of the data-base. If we are only interested in estimating functionals of $\mathbf{p}$ which do not depend on the identification of the different species with particular elements of $\mathcal{S}$, then we reduce the complexity of our statistical problem by reducing the data to $\mathbf{Y}$. The probability distribution of $\mathbf{Y}$ only depends on the reduced parameter $\mathbf{q}$. The possibility arises that for estimating functionals of $\mathbf{q}$, decreasing the complexity of the statistical problem by this (non-sufficient!) reduction of the data could be of benefit.

In the forensic statistical problem, we initially observe the database $\mathbf{X}$, which is then augmented by observing just one or two more animals. Prosecution and defence offer two different hypotheses to "explain" the additional data which has turned up in the specific crime case of interest. Under the prosecution hypothesis, the data-base is increased from size n to size $n+1$, leading to the observation of one animal of a previously unobserved species. Under the defence hypothesis, the data-base was increased from size n to $n+2$, leading to the observation of two different animals of the same but previously unobserved species.

Traditionally in forensic statistics, one is interested in the likelihood ratio defined as the ratio of the probabilities of the data arising in the crime under the two hypotheses of prosecution and defence. If the new species observed is species s, then these two probabilities are p_s and p_s^2, hence the likelihood ratio is $1/p_s$. However, the probability p_s is unknown, and earlier researchers have focussed attention on estimating p_s using the database, under a range of modelling assumptions and inference procedures. The problem of quantifying the uncertainty in the estimate has not been satisfactorily addressed, to date. Under our data reduction procedure, and treating the data (on the basis of which a likelihood ratio must be computed) as being database plus two alternative augmentations, *reduced* by "throwing away" the labels of the multinomial categories, i.e., the names of the species $s \in \mathcal{S}$, we obtain a new likelihood ratio equal to the ratio of the following two probabilities: (1), prosecution, the probability that when a sample is increased from size n to $n+1$, the new element belongs to a different species from all preceding observed; (2), defence, the probability that when a sample is increased from size n to $n+2$, the two new elements belongs to the same species which is different from all preceding observed.

Both of these two probabilities are functionals of the reduced parameter $\mathbf{q}$. They each can be estimated in a straightforward (but ad hoc) "Good-Turing" like way, using just the database. Alternatively, we can use the "plug-in" approach: estimate $\mathbf{q}$, using the database, in Orlitsky fashion (reduced data MLE) and then plug-in to obtain estimates of the functionals of interest. (Brenner estimates one of the probabilities in this way; for the other, he prefers a conservative approximation. We believe that in those cases where the Good-Turing estimator of the second probability is too unreliable, the Orlitsky approach still does work well. We see it as a modern alternative to a classical smoothing approach to "higher-order" Good-Turing estimators).

Initial experiments, and some theoretical work, suggests that we obtain in this way quite reliable estimates of a meaningful likelihood ratio. We are less ambitious, but, having "lowered our sights" and picked an easier target, we are now able to do a decent and, most importantly, *complete* job of statistical inference. Having estimated $\mathbf{q}$ in the Orlitsky way, we can use the parametric bootstrap to investigate, and report on, the precision of our "target functional", the base 10 logarithm of the reduced data likelihood ratio.

So far, Anevski et al. (2013) obtained (essentially) weak root-n consistency in L_1 norm of the reduced data MLE of $\mathbf{q}$. Our proof is inspired by a proof-outline given in one very short paper (a typical computer science conference proceedings submitted paper) by Orlitsky and his collaborators. It seems to us that the published outline proof is strictly speaking clearly wrong; more charitably, one could say that it is both very cryptic and clearly incomplete. However, this particular norm is not the one which really interests us. Moreover, the naive estimator has the same property, so this result does not indicate any superiority of the new estimator at all. At best, the result can be considered merely as a "sanity check". Numerical experimentation suggests that the reduced data MLE is *superior* to the "naive estimator" for estimating functionals of $\mathbf{q}$ which depend strongly on the tail of the distribution. Such estimation problems will also have lower than root-n rates of convergence. For a number of such functionals is it already *evident* that the naive estimator fails dramatically, while the Orlitsky estimator seems to do a rather good job - an impressively good job, in fact.

So on the theoretical side, we believe that we have stumbled across rather interesting statistical problems where a lot of further work can be done.

We also followed up suggestions of Orlitsky and his collaborators concerning computation of the reduced data MLE. The problem can be thought of as a missing data problem and the Expectation-Maximization (EM) algorithm comes naturally into view. However, the E step cannot be carried out explicitly in any but the smallest problems, and has to be replaced by a "stochastic" approximation. Here, Metropolis-Hastings (MH) seems to the only effective way to go about sampling from the required conditional distribution. The problem is a problem in combinatorial probability. However, "realistic" problems from our forensic science applied field are "large". EM is slow, and MH is slow; MH within EM is terribly slow. One can worry whether or not nice looking results are really due to the MLE being so

good, but are rather merely the result of starting a too slow algorithm at a "nice" starting point, and stopping long before convergence. Instead of nesting MH inside of EM we have experimented with a stochastic approximation (SA) approach in which MH and EM steps are carried out "in parallel". This gives some more confidence that the numerical solutions are the "good solutions" as well as giving a large range of user-choices for tuning the algorithm to faster convergence. Much more work needs to be done on these algorithms. Recent work of Piet Groeneboom has strongly increased our confidence in the numerical results we had obtained, and will probably give insight into how to improve our optimisation schedules.

Still, the "downside" of our approach is that the statistical estimation is definitely non-trivial and computer intensive. Here too, new ideas are needed.

References

[1] D. Anevski, R.D. Gill, S. Zohren, *Estimating a probability mass function with unknown labels*, submitted to *Ann. Statist.* arXiv:1312.1200 (2013).

Concentration inequalities for the exponential weighting method

Yuri Golubev

The talk deals with recovering an unknown vector $\mu \in \mathbb{R}^n$ from the noisy observations

$$Y_i = \mu_i + \sigma\,\xi_i, \quad i = 1, 2, \dots, n,$$

where ξ_i are independent Gaussian random variables with $\mathbf{E}\xi_i = 0$ and $\mathbf{E}\xi_i^2 = 1$. To simplify technical details, it assumed also that the noise level $\sigma > 0$ is known.

In what follows, vectors $(\mu_1, \dots, \mu_n)^\top$ and $(Y_1, \dots, \mu_n)^\top$ are denoted by μ and Y. The performance of an estimate $\hat{\mu}(Y) = (\hat{\mu}_1(Y), \dots, \hat{\mu}_n(Y))^\top$ is measured by l_2-losses $\|\hat{\mu}(Y) - \mu\|^2$, where $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$, i.e., $\|x\|^2 = \sum_{i=1}^n x_i^2$ and $\langle\cdot,\cdot\rangle$ stands for the standard inner product in $\mathbb{R}^n$.

The vector of interest μ is recovered with the help of the family of linear estimates

$$\hat{\mu}_i^h(Y) = h_i Y_i,\ i = 1, \dots, n, \quad h \in \mathcal{H},$$

where $\mathcal{H}$ is a given set of ordered multipliers (filters) [3] such that

- $h_i \in [0, 1],\ i = 1, \dots, n$ for all $h \in \mathcal{H}$,
- $h_{i+1} \le h_i,\ i = 1, \dots, n$ for all $h \in \mathcal{H}$,
- if for some integer k and some $h, g \in \mathcal{H}$, $h_k < g_k$, then $h_i \le g_i$ for all $i = 1, \dots, n$.

In order to construct a final estimate of μ, we compute a convex combination of $\hat{\mu}^h(Y)$, $h \in \mathcal{H}$, i.e.,

$$\bar{\mu}^w(Y) = \sum_{h \in \mathcal{H}} w^h(Y)\hat{\mu}^h(Y),$$

where weights $w^h(Y),\ h \in \mathcal{H}$ belong to the simplex

$$W = \Big\{ w^h \geq 0 : \sum_{h\in\mathcal{H}} w^h = 1 \Big\}.$$

This method goes back to [5] and often called convex aggregation. The main goal in this approach is to compute data-driven weights $w^h(Y)$ to minimize the losses $\|\mu - \bar{\mu}^w(Y)\|$ uniformly in $\mu \in \mathbb{R}^n$.

The standard solution to this problem is based on the unbiased risk estimation method which works as follows: let W' be a subset in W, compute

$$w(Y) = \arg\min_{w\in W'} \Big\{ \|Y - \bar{\mu}^w(Y)\|^2 + 2\sigma^2\langle \bar{h}^w, 1\rangle \Big\},$$

where $\bar{h}^w = \sum_{h\in\mathcal{H}} hw^h$.

Recall that the unbiased risk estimate of linear estimate $H \cdot Y$ is given by

$$\bar{r}(Y,H) = \|Y - H\cdot Y\|^2 + 2\sigma^2\langle H, 1\rangle.$$

This means that $\mathbf{E}\|\mu - H\cdot Y\|^2 = \mathbf{E}\bar{r}(Y,H) + n\sigma^2$.

Classical mathematical results in this approach are related to the model selection method [1]. Denote by $W_\circ = \big\{w \in W : w \in \{0,1\}\big\}$ the vertexes of W and compute

$$w_\circ(Y) = \arg\min_{w\in W_\circ} \Big\{ \|Y - \bar{\mu}^w(Y)\|^2 + 2\sigma^2\langle \bar{h}^w, 1\rangle \Big\}.$$

One can easily check that

$$w^h_\circ(Y) = \begin{cases} 1, & h = h^\circ(Y), \\ 0, & h \neq h^\circ(Y), \end{cases} \quad \text{where } h^\circ(Y) = \arg\min_{h\in\mathcal{H}} \Big\{ \|Y - h\cdot Y\|^2 + 2\sigma^2\langle h, 1\rangle \Big\}.$$

For the aggregated estimate

$$\bar{\mu}^{w_\circ}(Y) = h^\circ(Y)\cdot Y$$

the following theorem holds [3].

Theorem 1. *Uniformly in $\mu \in \mathbb{R}^n$*

$$\mathbf{E}\big\|\mu - \bar{\mu}^{w_\circ}(Y)\big\|^2 \leq r_\circ(\mu,\mathcal{H}) + C\sigma^2\sqrt{1 + \frac{r_\circ(\mu,\mathcal{H})}{\sigma^2}} \tag{1}$$

and

$$\mathbf{P}\Big\{ \big\|\mu - \bar{\mu}^{w_\circ}(Y)\big\| \geq \sqrt{r_\circ(\mu,\mathcal{H})} + \sigma x \Big\} \leq \exp(-Cx^2), \tag{2}$$

where $r_\circ(\mu,\mathcal{H}) = \min\limits_{w\in W_\circ} \mathbf{E}\|\bar{\mu}^w(Y) - \mu\|^2$ is the so-called oracle risk and C is a universal constant.

The main results in this talk concern a generalization of the model selection approach called *Exponential Weighting* (EW). Let $\pi \in W$ be a probability distribution on $\mathcal{H}$ (a priory weights). The motivation of the EW is based on the following ideas :

- to upper-bound the unbiased risk estimate $\bar{r}(Y, \bar{h}^w)$ by

$$\bar{r}(Y, \bar{h}^w) \leq \sum_{h \in \mathcal{H}} w^h \bar{r}(Y, h);$$

- to control the Kulback-Leibler divergence between $w(Y)$ and π

$$K(\pi, w) = \sum_{h \in \mathcal{H}} \pi^h \log \frac{\pi^h}{w^h}.$$

Therefore the exponential weights are defined by

$$w_\beta(Y) = \arg\min_{w \in W} \left\{ \sum_{h \in \mathcal{H}} w^h \bar{r}(Y, h) + 2\sigma^2 \beta K(\pi, w) \right\},$$

where $\beta \geq 0$ is often called temperature. With a simple algebra one obtains

$$w_\beta^h(Y) = \pi^h \exp\left[-\frac{\bar{r}(Y, h)}{2\beta\sigma^2} \right] \left\{ \sum_{h' \in \mathcal{H}} \pi^{h'} \exp\left[-\frac{\bar{r}(Y, h')}{2\beta\sigma^2} \right] \right\}^{-1}.$$

In order to mimic the oracle risk $r_\circ(\mu, \mathcal{H})$ with the help of the aggregated estimate

$$\bar{\mu}_\beta(Y) = \sum_{h \in \mathcal{H}} w_\beta^h(Y)\, \mu^h(Y),$$

the following conditions are assumed :

- $$\pi^h \stackrel{\text{def}}{=} 1 - \exp\left\{ -\frac{\langle h^+, 1\rangle - \langle h, 1\rangle}{\beta} \right\}, \text{ where } h^+ = \min\{g \in \mathcal{H} : g > h\}$$

  and $\pi^{h_{\max}} = 1$, where $h^{\max}$ is the maximal multiplier in $\mathcal{H}$.
- There exist constants $K_\circ$ and K° such that

$$\|h\|^2 - \|g\|^2 \geq K_\circ\big(\langle h, 1\rangle - \langle g, 1\rangle\big) \quad \text{for all } h \geq g,$$
$$\|h^+\|^2 \leq K^\circ \|h\|^2 \quad \text{for all } h \in \mathcal{H}.$$

The following theorem controlling the concentration of $\big\|\mu - \bar{\mu}_\beta(Y)\big\|$ is the main result in this talk.

Theorem 2. *If $\beta \geq 4$, then uniformly in $\mu \in \mathbb{R}^n$*

$$\mathbf{E}\big\|\mu - \bar{\mu}_\beta(Y)\big\|^2 \leq r_\beta(\mu, \mathcal{H}) + C\sigma^2; \tag{3}$$

for any $\beta > 0$, uniformly $\mu \in \mathbb{R}^n$

$$\mathbf{P}\Big\{ \big\|\mu - \bar{\mu}_\beta(Y)\big\| \geq \sqrt{r_\beta(\mu, \mathcal{H})} + \sigma x \Big\} \leq \exp\big[-Cx^2\big], \tag{4}$$

where $C = C(\beta, K_\circ, K^\circ)$ is a constant and

$$r_\beta(\mu, \mathcal{H}) = r_\circ(\mu, \mathcal{H}) + 2\beta\sigma^2 \log\left[1 + \frac{r_\circ(\mu, \mathcal{H})}{\sigma^2} \right].$$

The proof of Theorem 2 is essentially based on [4] and [2]. Notice also that Equation (3) improves substantially the upper bound in (1), whereas the concentration inequalities for the model selection method (2) and for the EW method (4) are almost equivalent.

References

[1] H. Akaike, *Information theory and an extension of the maximum likelihood principle* , Proc. 2-nd Intern. Symp. Inf. Theory (1973), 267–281.

[2] E. Chernousova, Yu. Golubev, and E. Krymova, *Ordered Smoothers With Exponential Weighting*, Electronic J. Statist. **7** (2013), 2395–2419.

[3] A. Kneip, *Ordered linear smoothers*, Annals of Statist. **22** (1994), 835–866.

[4] G. Leung and A. Barron, *Information theory and mixing least-squares regressions*, IEEE Transactions on Information Theory **52** (2006), 3396–3410.

[5] A. Nemirovski, A. *Topics in non-parametric statistics*, Lectures Notes in Math. **1738** (2000), Springer-Verlag, Berlin.

Conditional limit theorems for products of random matrices

Ion Grama

(joint work with Emile Le Page, Marc Peigné)

Let $\mathbb{G} = GL(d, \mathbb{R})$ be the general linear group of $d \times d$ invertible matrices w.r.t. ordinary matrix multiplication. If g is an element of of $\mathbb{G}$ by $\|g\|$ we mean the operator norm and if v is an element of the vector space $\mathbb{V} = \mathbb{R}^d$ the norm $\|v\|$ is Euclidean. Endow the group $\mathbb{G}$ by the usual Borel σ-algebra w.r.t. $\|\cdot\|$. Let $\boldsymbol{\mu}$ be a probability measure on $\mathbb{G}$ and suppose that on the probability space $(\Omega, \mathcal{F}, \mathbf{Pr})$ we are given an i.i.d. sequence $(g_n)_{n\geq 1}$ of $\mathbb{G}$-valued random elements of the same law $\mathbf{Pr}(g_1 \in dg) = \boldsymbol{\mu}(dg)$. A random walk in $\mathbb{G}$ is the product $G_n = g_n \dots g_1$. Let $v \in \mathbb{V} \setminus \{0\}$ be a any starting point. The object of interest is the size of the vector $G_n v$ which is controlled by the quantity $\log \|G_n v\|$. It follows from the results of Le Page [3] that, under appropriate assumptions, the sequence $(\log \|G_n v\|)_{n\geq 1}$ behaves like a sum of i.i.d. r.v.'s and satisfies standard classical properties such as the law of large numbers, law of iterated logarithm and the central limit theorem. There is a vaste literature on this subject. We refer to Bougerol and Lacroix [1] and to the references therein.

Introduce the following conditions. Let $N(g) = \max\left\{\|g\|, \|g\|^{-1}\right\}$, $\mathrm{supp}\boldsymbol{\mu}$ be the support of the measure $\boldsymbol{\mu}$ and $\mathbb{P}(\mathbb{V})$ be the projective space of $\mathbb{V}$.

P1. *There exists* $\delta_0 > 0$ *such that*

$$\int_{\mathbb{G}} N(g)^{\delta_0} \boldsymbol{\mu}(dg) < \infty,$$

The next condition requires, roughly speaking, that the dimension of the support of $\mathrm{supp}\boldsymbol{\mu}$ cannot be reduced.

P2 (Strong irreducibility). *The support* $\text{supp}\boldsymbol{\mu}$ *of* $\boldsymbol{\mu}$ *acts strongly irreducibly on* $\mathbb{V}$, *i.e. no proper union of finite vector subspaces of* $\mathbb{V}$ *is invariant with respect to all elements* g *of the group generated by* $\text{supp}\boldsymbol{\mu}$.

We say that the sequence $(h_n)_{n\geq 1}$ of elements of $\mathbb{G}$ is contracting for the projective space $\mathbb{P}(\mathbb{V})$ if $\lim_{n\to\infty} \log \frac{a_1(n)}{a_2(n)} = \infty$, where $a_1(n) \geq ... \geq a_d(n)$ are the eigenvalues of the symmetric matrix $h_n' h_n$ and h_n' is the transpose of h_n.

P3 (Proximality). *The closed semigroup generated by* $\text{supp}\boldsymbol{\mu}$ *contains a contracting sequence for the projective space* $\mathbb{P}(\mathbb{V})$.

In the sequel for any $v \in \mathbb{V} \setminus \{0\}$ we denote by $\overline{v} = \mathbb{R}v \in \mathbb{P}(\mathbb{V})$ its direction and for any direction $\overline{v} \in \mathbb{P}(\mathbb{V})$ we denote by v a vector in $\mathbb{V} \setminus \{0\}$ of direction $\overline{v}$. Define the function $\rho : \mathbb{G} \times \mathbb{P}(\mathbb{V}) \to \mathbb{R}$ called norm cocycle by setting

$$\rho(g, \overline{v}) := \log \frac{\|gv\|}{\|v\|}, \text{ for } (g, \overline{v}) \in \mathbb{G} \times \mathbb{P}(\mathbb{V}).$$

It is well known (see Le Page [3] and Bougerol and Lacroix [1]) that under conditions **P1-P3** there exists an unique $\boldsymbol{\mu}$-invariant measure $\boldsymbol{\nu}$ on $\mathbb{P}(\mathbb{V})$ such that, for any continuous function φ on $\mathbb{P}(\mathbb{V})$,

$$(\boldsymbol{\mu} * \boldsymbol{\nu})(\varphi) = \boldsymbol{\nu}(\varphi).$$

Moreover the upper Lyapunov exponent

$$\gamma = \gamma_{\boldsymbol{\mu}} = \int_{\mathbb{G}\times\mathbb{P}(\mathbb{V})} \rho(g, \overline{v})\, \boldsymbol{\mu}(dg)\, \boldsymbol{\nu}(d\overline{v})$$

is finite and there exists a constant $\sigma > 0$ such that for any $v \in \mathbb{V} \setminus \{0\}$ and any $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbf{Pr}\left(\frac{\log \|G_n v\| - n\gamma}{\sigma\sqrt{n}} \leq t\right) = \Phi(t),$$

where $\Phi(\cdot)$ is the standard normal distribution.

Denote by $\mathbb{B}$ the closed unit ball in $\mathbb{V}$ and by $\mathbb{B}^c$ its complement. For any $v \in \mathbb{B}^c$ define the exit time of the random process $G_n v$ from $\mathbb{B}^c$ by

$$\tau_v = \min\{n \geq 1 : G_n v \in \mathbb{B}\}.$$

In the sequel we consider that the upper Lyapunov exponent γ is equal to 0. The fact that $\gamma = 0$ does not imply that the events

$$\{\tau_v > n\} = \{G_k v \in \mathbb{B}^c : k = 1, ..., n\},\ n \geq 1$$

occur with positive probability for any $v \in \mathbb{B}^c$. To ensure this we need the following additional condition:

P4. *There exists* $\delta > 0$ *such that*

$$\inf_{s\in\mathbb{S}^{d-1}} \boldsymbol{\mu}(g : \log \|gs\| > \delta) > 0.$$

Under conditions **P1-P4** we prove that, for any $v \in \mathbb{B}^c$,

$$\mathbf{Pr}(\tau_v > n) = \frac{2V(v)}{\sigma\sqrt{2\pi n}}(1 + o(1)) \text{ as } n \to \infty,$$

where V is a positive function on $\mathbb{B}^c$. Moreover, we prove that the limit law of the quantity $\frac{1}{\sigma\sqrt{n}} \log \|G_n v\|$, given the event $\{\tau_v > n\}$ coincides with the Rayleigh distribution $\Phi^+(t) = 1 - \exp\left(-\frac{t^2}{2}\right)$: for any $v \in \mathbb{B}^c$ and for any $t \geq 0$,

$$\lim_{n\to\infty} \mathbf{Pr}\left(\frac{\log \|G_n v\|}{\sigma\sqrt{n}} \leq t \middle| \tau_v > n \right) = \Phi^+(t).$$

Our proofs rely upon a strong approximation result for Markov chains established in [2].

References

[1] Bougerol, P. and Lacroix J. *Products of Random Matrices with Applications to Schödinger Operators.* (1985), Birghäuser, Boston-Basel-Stuttgart.

[2] Grama, I., Le Page, E. and Peigné, M. *On the rate of convergence in the weak invariance principle for dependent random variables with applications to Markov chains.* Colloq. Math. 134 (2014), 1-55.

[3] Le Page, E. *Théorèmes limites pour les produits de matrices aléatoires. Springer Lecture Notes,* **928**, (1982) 258-303.

Nuclear-norm regularization for quantum and classical estimation problems

David Gross

The theory of compressed sensing provides rigorous methods for analyzing the performance of estimators that include a sparsity-enhancing ℓ_1-norm regularization term. Since around 2009, a "non-commutative" version of compressed sensing has been developed [1, 2, 3]. Here, the aim is to efficiently recover matrices under a low-rank assumption, most commonly using nuclear-norm regularization. The program was initially motivated by purely classical estimation problems – e.g. the influential "Netflix problem" of predicting user preferences in online shops. However, early on, a fruitful interaction between classical and quantum theory ensued: In one direction, it has been realized that low-rank methods lead to rigorous and very tight performance guarantees for quantum state estimation procedures [4, 5]. In the other direction, mathematical methods originally developed in the context of quantum information theory allowed for a significant generalization and simplification of the rigorous results on low-rank recovery [3, 6].

In this extended abstract, I will focus on one particular aspect of my talk: the problem of how to efficiently construct tight confidence regions for estimates of quantum states. This problem featured heavily in discussions during the workshop, was part of the "open problems" session, and seems to be a very fruitful subject for collaborations between mathematical statisticians and quantum physicists. My aim is to briefly sketch the mathematical problem, without going into any detail or mentioning any of the underlying physics.

Every quantum system is associated with a dimension d and a d-dimensional Hilbert space $\mathcal{H}$. Mathematically, its *state space* is the set $\mathcal{S}(\mathcal{H})$ of positive semi-definite (psd), unit-trace operators on $\mathcal{H}$. They parameterize the statistical model. To make things more concrete, we assume that a basis of $\mathcal{H}$ has been chosen. Pick a particular state $\rho \in \mathcal{S}(\mathcal{H})$ and denote its matrix elements w.r.t. to said basis by $\theta \in \mathbb{R}^p$, with $p := d^2$. Physical measurements are described by a standard linear model

$$Y = X\theta + \epsilon, \tag{1}$$

where the design matrix $X \in \mathbb{R}^{n\times p}$ depends on the physical procedure used to perform the measurements, $Y \in \mathbb{R}^n$ are the observed quantities, and ϵ is a mean-zero variable describing statistical noise. Quantum mechanics demands that X fulfill a set of non-trivial positivity constraints, but we are going to ignore these details here (see e.g. [7]).

The *quantum state estimation problem* now asks for the construction of estimators $\hat{\theta}(Y)$ and associated confidence regions $\hat{C}(Y)$. The non-trivial aspect here is to incorporate the psd and unit-trace constraint on the ρ. Practitioners usually employ either Bayesian methods [8] or numerical maximum-likelihood estimators [9] for $\hat{\theta}$, together with bootstrap-type methods for finding confidence regions $\hat{C}$.

There is a variety of reasons not to be satisfied with the state of the art. First, it has been argued [10] that ML-based estimators of $\hat{C}$ tend to be optimistic. Thus, procedures that would produce fairly tight confidence regions in an computationally efficient way, while being endowed with mathematical sound statistical guarantees, would be highly desirable. Second, as mentioned in the introduction, ideas introduced to the quantum estimation problem from the fields of compressed sensing, low-rank recovery and nuclear-norm regularizations, showed that a fairly non-trivial structure exists. We believe this makes the problem interesting also from a purely theoretical perspective.

Indeed, the series of papers [4, 3, 6, 5] has established recovery guarantees for ρ of known rank, with a particular focus on non-invertible design matrices X (i.e. $n < p$). (This program led to a series of new results on non-uniform [3] and uniform [6] low-rank matrix recovery). The papers [4, 5] also gave first constructions for *adaptive* region estimators. These are region estimators $\hat{C}(Y)$ that come close to achieving the mini-max risk for every given value of the rank r, even if the rank is unknown. The details of these results are somewhat difficult to compare to related statements appearing in the mathematical statistics literature (in particular [11]). The first reason is that the authors of [4, 5] were, at the time, unaware of the existing framework for phrasing such statements and developed them in an *ad hoc* manner. The second reason is that both papers also incorporate to varying degree certain positivity and locality constraints on the design matrix X. Explaining either constriant in detail is beyond the scope of this extended abstract. To anyway give a rough idea of the results: Both the scheme in [4] and in [5] depend on some form of *sample splitting*. In [4], the idea is to first estimate the rank r from parts of the data, and then estimate $\hat{\theta}$ and $\hat{C}$ making

use of that information. The approach of [5] is somewhat complementary. There, it is proposed to first constructs an estimate $\hat{\theta}$ and then employ a scheme called *direct fidelity estimation* [12] to construct $\hat{C}$. An issue not completely addressed in [12] is that the regularization parameter used in the first step already depends on the rank. In practice, one would use a generous guess for r to obtain $\hat{\theta}$ and rely on the fact that the fidelity estimation procedure for $\hat{C}$ gives results that do not depend on how $\hat{\theta}$ has been constructed. While this does allow for constructing rigorously justified confidence regions, the global protocol is somewhat ill-defined, as no rigorous guidance is offered on how to choose the "generous guess for r" in the first place.

In the model employed in [5], each row of the design matrix X represents a physical experiment, which one is generally free to design as one pleases. In particular, one can decrease the noise (measured e.g. in terms of $\mathrm{Var}(\epsilon_i)$), by repeating any given experiment a chosen number k_i of times. Thus the following setup becomes natural: Assume that each row of the design matrix X is chosen independently and uniformly from a fixed orthogonal basis of the set of $(d \times d)$-matrices (e.g. the *Pauli basis* [4]). Every such row X_i translates into a quantum experiment that gives rise to a Bernoulli random variable whose expectation value is proportional to $(X\theta)_i$ (the precise relation depends on the normalization convention). We will repeat each experiment k times for some number k to be specified later. The variable Y_i is the average over these k experiments. Now assume that the true ρ has rank at most r and that X has $n = Crd \log^6 d$ rows, selected uniformly from the Pauli basis. The question treated in [5] is: How big does k have to be in order for the risk (measured in nuclear norm) of the nuclear-norm regularized *Dantzig selector* estimator to reach a certain prescribed level ϵ? The answer is $k = O\big(\frac{rd}{\epsilon}\big)$. The paper goes on to describe how to construct a confidence region $\hat{C}$. It will be a ball of radius ϵ around $\hat{\theta}$, with the radius measured in terms of the *fidelity* – a measure broadly equivalent to nuclear norm. In order to achieve it, an additional $O\big(\frac{\hat{r}^4}{\epsilon^4} d \log \hat{r}\big)$ Bernoulli experiments have to be performed (this statement is greatly simplified – c.f. [5] for details). Here, $\hat{r}$ is the rank of the *estimate*, i.e. a known quantity.

These results show that non-trivial confidence regions for quantum state estimation can be constructed. In principle, their diameter decreases with the unknown rank. However, a host of open problems remain: (1) A rigorous protocol that depends at no point on a rough estimate for the rank has not yet been given (even though the coverage of the confidence regions of [5] does not depend on that guess). (2) The analysis seems far from tight. (3) The "sample splitting" approach feels artificial. (4) There are low-dimensional models other than low-rank which, in principle, allow for improved confidence regions. Examples are quantum versions of hidden Markov models [13, 14, 15]. Protocols giving adaptive confidence sets for those models from physically realistic measurements are still missing.

It is our hope that the discussions initiated during the workshop will lead to the resolution of some of these questions, possibly along the lines of [11].

References

[1] B. Recht, M. Fazel, PA. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM Review **52** (2010), 471–501.
[2] E. Candès, T. Tao, *The power of convex relaxation: Near-optimal matrix completion*, IEEE Transactions on Information Theory, **56** (2010), 2053–2080.
[3] D. Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Transactions on Information Theory, **57** (2011), 1548–1566.
[4] D. Gross, Y.-K. Liu, S. Flammia, S. Becker, J. Eisert, *Quantum state tomography via compressed sensing*, Phys. Rev. Lett. **105** (2010), 150401.
[5] S. Flammia, D. Gross, Y.-K. Liu, J. Eisert, *Quantum Tomography via Compressed Sensing: Error Bounds, Sample Complexity, and Efficient Estimators* New J. Phys. **14** (2012), 095022.
[6] Y.-K. Liu, *Universal low-rank matrix recovery from Pauli measurements*, Advances in Neural Information Processing Systems (NIPS) **24** (20122), 1638.
[7] D. Petz, *Quantum Information Theory and Quantum Statistics*, Springer 2007.
[8] K. Audenaert, S. Scheel, *Quantum tomographic reconstruction with error bars: a Kalman filter approach*, New Journal of Physics **11** (2009), 023028.
[9] M. Paris, J. Rehacek (edts.), *Quantum State Estimation*, Springer 2004.
[10] R. Blume-Kohout, *Hedged maximum likelihood estimation*, Phys. Rev. Lett. **105** (2010), 200504.
[11] R. Nickl, S. van de Geer, *Confidence sets in sparse regression*, The Annals of Statistics **41** (2013), 2852-2876.
[12] S. Flammia, Y.K. Liu, *Direct fidelity estimation from few Pauli measurements*, Physical Review Letters **106** (2011), 230501.
[13] M. Cramer, *et al.*, *Efficient quantum state tomography*, Nature Communications **1** (2010), 149.
[14] T. Baumgratz, D. Gross, M. Cramer, M. Plenio, *Scalable reconstruction of density matrices*, Phys. Rev. Lett. **111** (2013), 020401.
[15] M. Guta, J. Kiukas, *Equivalence classes and local asymptotic normality in system identification for quantum Markov chains*, pre-print arXiv:1402.3535 (2014).

Quantum Gaussian optimizers problem

Alexander S. Kholevo (Holevo)
(joint work with Vittorio Giovannetti, Andrea Mari)

Proving the quantum Gaussian optimizers conjecture is an analytical optimization problem that arose in quantum information theory some time ago. Its difficulty can be seen from analogy with the related classical problem of Gaussian maximizers which has been studied rather exhaustively, see Lieb [6] and references therein. Consider an integral operator G from $L_p(\mathbb{R}^s)$ to $L_q(\mathbb{R}^s)$ given by a Gaussian kernels with the $(q \to p) -$ norm

$$\|G\|_{q\to p} = \sup_{\|f\|_q \le 1} \|Gf\|_p \,. \tag{1}$$

Under rather general circumstances this operator is correctly defined, and the supremum in (1) is attained on a Gaussian function f. Moreover, under some additional restrictions any maximizer is Gaussian. A difficulty in the optimization problem (1) is that it requires *maximization* of a convex function, so the general

theory of convex optimization is not of great use here. Instead, the solution is heavily based on a classical Minkowski's inequality and the related multipicativity of the classical $(q \to p)$ −norms with respect to tensor products of the integral operators.

The quantum Gaussian optimizers problem refers to Bosonic Gaussian channels – a noncommutative analog of Gaussian kernels and, similarly, requires maximization of convex functions (or minimization of concave functions, such as entropy) of the output state of the channel. A general conjecture appeared in [5] (see also [3]) is that the optimizers belong to the class of pure Bosonic Gaussian states. The conjecture, however natural it looks, resisted numerous attacks for a dozen of years, for a survey see e.g. [8], [1]. There could be a hope that in solving the problem one could use the classical " Gaussian maximizers" results. However the noncommutative analog of the Minkowski's inequality is not powerful enough to guarantee the multipicativity of norms (or additivity of the corresponding entropic quantities). Moreover, the related long-standing multipicativity/additivity problem in quantum information theory was recently shown to have negative solution in general by Winter, Hayden and Hastings, see [3] for a survey. Instead, a solution of the quantum Gaussian optimizers conjecture found in [1], [7] uses completely different ideas based on a thorough study of structural properties of quantum Gaussian channels. Remarkably, the present solution of the multidimensional quantum Gaussian optimizers problem [2] implies also a proof of the multipicativity/additivity property in the restricted class of gauge-covariant or contravariant quantum Gaussian channels.

It would be interesting to investigate a possible development of such an approach to obtain noncommutative generalizations of the classical " Gaussian maximizers" results for $(q \to p)$ −norms. Such a generalization could shed a new light to the hypercontractivity problem for quantum dynamical semigroups and related noncommutative analogs of logarithmic Sobolev inequalities, see [9] for the case of finite-level quantum systems.

The solution of the conjecture is restricted to channels that are *gauge-covariant* or *contravariant* with respect to fixed complex structure. Therefore we consider the classical phase space $\mathbf{Z}$ consisting of $s-$dimensional complex column vectors $\mathbf{z}$. The gauge group acts in $\mathbf{Z}$ as multiplication by $e^{i\phi}$, where ϕ is phase and the complex-linear operators in $\mathbf{Z}$ are represented by complex $s \times s-$matrices. The quantized system is described by the displacement operators $D(\mathbf{z}) = \exp\left(\mathbf{a}^{\dagger}\mathbf{z} - \mathbf{z}^{*}\mathbf{a}\right)$, where $\mathbf{a}$ is the vector of annihilation operators for s Bosonic modes in the system Hilbert space.

The action of a Gaussian gauge-covariant channel in the Heiseberg picture can be described as

$$\Phi^{*}[D(\mathbf{z})] = D(\mathbf{K}\mathbf{z}) \exp\left(-\mathbf{z}^{*}\boldsymbol{\mu}\boldsymbol{z}\right), \quad \mathbf{z} \in \mathbf{Z}, \tag{2}$$

where $\mathbf{K}$ is complex $s \times s-$matrix, $\boldsymbol{\mu}$ is Hermitian $s \times s-$matrix satisfying the condition (see [3])

$$\boldsymbol{\mu} \geq \pm\frac{1}{2}\left(\mathbf{I} - \mathbf{K}^{*}\mathbf{K}\right), \tag{3}$$

where $\mathbf{I}$ is the unit $s \times s-$matrix.

Denoting by $\bar{\mathbf{z}}$ the column vector obtained by taking the complex conjugate of the elements of $\mathbf{z}$, the action of the Gaussian gauge-contravariant channel is described as

$$\Phi^*[D(\mathbf{z})] = D(-\overline{\mathbf{Kz}}) \exp\left(-\mathbf{z}^* \boldsymbol{\mu} \boldsymbol{z}\right), \tag{4}$$

where $\boldsymbol{\mu}$ is Hermitian matrix satisfying the inequality

$$\boldsymbol{\mu} \geq \frac{1}{2}\left(\mathbf{I} + \mathbf{K}^*\mathbf{K}\right), \tag{5}$$

Theorem. (i) *Let Φ be a gauge covariant or contravariant channel and let f be a real concave function on $[0,1]$, such that $f(0) = 0$, then*

$$\mathrm{Tr} f(\Phi[\rho]) \geq \mathrm{Tr} f(\Phi[|\mathbf{w}\rangle\langle\mathbf{w}|]) = \mathrm{Tr} f(\Phi[|\mathbf{0}\rangle\langle\mathbf{0}|]) \tag{6}$$

for all states ρ and any coherent state $|\mathbf{w}\rangle\langle\mathbf{w}|$ (the value on the right is the same for all coherent states by the unitary covariance property of a Gaussian channel.

(ii) *Let f be strictly concave, $\mathbf{K}\mathbf{K}^* > 0$, and (3), (5) hold as strict inequalities, then the equality in (6) is attained only if ρ is a coherent state.*

By taking $f(x) = -x^p$, $f(x) = -x \log x$, we obtain that the $(1 \to p)$-norm $\|\Phi\|_{1\to p}$, the minimal Rényi entropy $\check{R}_p(\Phi)$ and the minimal von Neumann entropy $\check{H}(\Phi)$ of the channel Φ are all optimized by the input vacuum state $|0\rangle\langle 0|$. From the definitions of gauge-co/contravariant channels (2), (4), it follows that the state $\Phi[|0\rangle\langle 0|]$ is gauge-invariant Gaussian with the correlation matrix $\boldsymbol{\mu} + \mathbf{K}^*\mathbf{K}/2$. The spectrum of $\Phi[|0\rangle\langle 0|]$ is computed explicitly [4] leading to the explicit expressions for these quantities [2].

If Φ_1 and Φ_2 are both gauge-covariant (contravariant), then their tensor product $\Phi_1 \otimes \Phi_2$ shares the same property. The multiplicativity property of $(1 \to p)$-norms for any two Gaussian gauge-covariant (contravariant) channels Φ_1 and Φ_2, as well as the additivity of the minimal Rényi entropies and of the minimal von Neumann entropy then follows from the product property of the optimizing vacuum state.

References

[1] V. Giovannetti, A. S. Holevo, R. Garcia-Patron, *A solution of the Gaussian optimizer conjecture*, arXiv:1312.2251, 2013.

[2] V. Giovannetti, A. S. Holevo, A. Mari, *Majorization and additivity for multimode bosonic Gaussian channels*, arXiv:1405.4066, 2014.

[3] A. S. Holevo, *Quantum systems, channels, information. A mathematical introduction*, De Gruyter, 2012.

[4] A. S. Holevo, M. Sohma, O. Hirota, *Error exponents for quantum channels with constrained inputs*, Rep. Math. Phys., **46** (2000), 343–358.

[5] A. S. Holevo, R. A. Werner. *Evaluating capacities of Bosonic Gaussian channels*, Phys. Rev. A, 63:032312, 2001.

[6] E. H. Lieb, *Gaussian kernels have only Gaussian maximizers*, Invent. Math., 102:179-208, 1990.

[7] A. Mari, V. Giovannetti, A. S. Holevo, *Quantum state majorization at the output of bosonic Gaussian channels*, Nature Communications, 5:3826, 2014.

[8] A. Serafini, J. Eisert, M. M. Wolf, *Multiplicativity of maximal output purities of Gaussian channels under Gaussian inputs*, Phys. Rev. A, 71:012320, 2005.

[9] K. Temme, F. Pastawski, M. J. Kastoryano. *Hypercontractivity of quasi-free quantum semigroups*, arXiv:1403.5224, 2014.

Statistical inference for high-dimensional estimation of the inverse covariance matrix

Jana Janková
(joint work with Sara van de Geer)

We propose methodology for estimation of large sparse precision matrices and statistical inference for their low-dimensional parameters in a high-dimensional setting where the number of parameters can be much larger than the sample size. Many procedures for estimation of precision matrices have been proposed, which are typically based on thresholding and hence lead to sparse estimators whose asymptotic distribution depends on the true unknown parameter and the convergence to the limit is not uniform. This poses a major difficulty in developing methodology for quantifying the uncertainty of estimation.

We propose a general way of modifying a sparse estimator to obtain an estimator of the precision matrix whose entries have a Gaussian limiting distribution. This results in a way of constructing confidence regions and hypothesis testing for low-dimensional parameters. The proposed de-sparsified estimator enjoys rate optimality in supremum norm over a large class of sparse precision matrices and thresholding it gives guarantees for variable selection. Similar approach has been adopted in the context of high-dimensional linear regression [6], [5]. This line of work is inspired by the semi-parametric theory [1], where one concentrates on a low-dimensional parameter of interest and considers the high-dimensional part as a nuisance parameter.

Suppose that we observe an i.i.d. sample $X_1, \ldots, X_n \sim \mathcal{N}_p(0, \Sigma_0)$, where the covariance matrix $\Sigma_0 \in \mathbb{R}^{p\times p}$ is unknown. The goal is to estimate the precision matrix $\Theta_0 := \Sigma_0^{-1}$ (assuming the inverse of Σ_0 exists) in a setting where $p \gg n$ and Θ_0 is a (sufficiently) sparse matrix with spectrum bounded uniformly in n. We remark here that the Gaussianity assumption on the design $X_1, \ldots, X_n$ is in fact not necessary to establish asymptotic normality and may be relaxed to a sub-Gaussian tail assumption on the margins of the underlying distribution as discussed below.

For a constant $L \geq 1$ and a sequence s_n we define the model

$$\begin{aligned}\mathcal{G}_n(s_n, L) \quad := \quad & \{\Theta \in \mathbb{R}^{p\times p} : \max_{i=1,\ldots,p} |\{j : \Theta_{ij} \neq 0\}| \leq s_n, \\ & 1/L \leq \Lambda_{\min}(\Theta) \leq \Lambda_{\max}(\Theta) \leq L\}.\end{aligned}$$

Here s_n, the upper bound on row sparsity of Θ_0, depends on n, and appropriate restrictions on s_n are presented separately.

To illustrate the procedure to obtain an asymptotically normal estimator of the precision matrix, consider the nodewise regression estimator [3] with tuning parameters $\lambda_j = \lambda \asymp \sqrt{\log p/n}, j = 1, \ldots, p$. The nodewise regression estimator $\hat{\Theta}$

is characterized by the Karush-Kuhn-Tucker (KKT) conditions

$$\hat{\Sigma}\hat{\Theta} - I + \lambda\hat{Z} = 0,$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$ is the sample covariance matrix and $\hat{Z}$ is a term arising from the sub-differential of the ℓ_1 penalty. We "invert" the KKT conditions with an approximate inverse $\hat{\Theta}$ of the sample covariance matrix $\hat{\Sigma}$ to obtain

$$\hat{\Theta}^T\hat{\Sigma}\hat{\Theta} - \hat{\Theta}^T + \hat{\Theta}^T\lambda\hat{Z} = 0.$$

Rearranging, one can show under $s_n = o(n/\log p)$ and using the ℓ_1 rates of convergence of the nodewise regression estimator,

$$\hat{\Theta} - \hat{\Theta}^T\lambda\hat{Z} - \Theta_0 = \Theta_0(\hat{\Sigma} - \Sigma_0)\Theta_0 + o_{\mathbb{P}}\left(s_n \log p/n\right), \tag{1}$$

where $\max_{i=1,\ldots,p} |\{j : \Theta^0_{ij} \neq 0\}| \leq s_n$. Consequently, under sparsity $s_n = o\left(\sqrt{n}/\log p\right)$, the remainder term in (1) is of order $n^{-1/2}$ and the leading term is asymptotically normal. This suggests to take the de-biased estimator

$$\hat{T} := \hat{\Theta} - \hat{\Theta}^T\lambda\hat{Z} = \hat{\Theta} + \hat{\Theta}^T - \hat{\Theta}\hat{\Sigma}\hat{\Theta}, \tag{2}$$

as an estimator of Θ_0.

Theorem 1. *Suppose that $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(0, \Sigma_0)$, where $\Theta_0 := \Sigma_0^{-1}$ exists, $\Theta_0 \in \mathcal{G}_n(s_n, L)$ where $L = \mathcal{O}(1)$ and $s_n = o(\sqrt{n}/\log p)$. Let $\hat{T}$ be the estimator defined in (2) with $\lambda_j \asymp \sqrt{\log p/n}$ uniformly in $j = 1, \ldots, p$. Then for every $(i, j) \in \{1, \ldots, p\}^2$ and $z \in \mathbb{R}$ it holds*

$$\lim_{n\to\infty} \sup_{\Theta_0 \in \mathcal{G}_n(s_n, L)} |\mathbb{P}_{\Theta_0}\left(\sqrt{n}(\hat{T}_{ij} - \Theta^0_{ij})/\hat{\sigma}_{ij} \leq z\right) - \Phi(z)| = 0,$$

where $\hat{\sigma}^2_{ij} := \hat{\Theta}_{ii}\hat{\Theta}_{jj} + \hat{\Theta}^2_{ij}$.

Consequently, Theorem 1 establishes uniform convergence of the entries of $\hat{T}$ to the Gaussian distribution and thus leads to uniformly (over $\mathcal{G}_n(s_n, L)$) valid confidence regions for low-dimensional parameters of the precision matrix.
Furthermore, when $s_n = o(\sqrt{n}/\log p)$ one can show that the proposed estimator achieves the rate $n^{-1/2}$ for individual entries of $\hat{T}$ (and $\sqrt{\log p/n}$ for the maximum of all entries). When only $s_n = o(n/\log p)$, the estimator achieves the minimax rate bound in supremum norm [4] and hence is in this sense optimal as shown in Theorem 2 below.

Theorem 2. *Suppose that $X_1, \ldots, X_n$ are i.i.d. $\mathcal{N}(0, \Sigma_0)$, where $\Theta_0 := \Sigma_0^{-1}$ exists, $\Theta_0 \in \mathcal{G}_n(s_n, L)$ where $L = \mathcal{O}(1)$ and $s_n = o(n/\log p)$. Let $\hat{T}$ be the estimator defined in (2) with $\lambda_j \asymp \sqrt{\log p/n}$ uniformly in $j = 1, \ldots, p$. Then for all $\varepsilon > 0$ there exists $C_\varepsilon > 0$*

$$\sup_{\Theta_0 \in \mathcal{G}(s_n, L)} \mathbb{P}_{\Theta_0}\left(\|\hat{T} - \Theta^0\|_\infty > C_\varepsilon \max\left\{\sqrt{\log p/n}, s\log p/n\right\}\right) < \varepsilon.$$

The non-sparse estimator $\hat{T}$ may be thresholded to obtain exact recovery of the active set where the coefficients are sufficiently larger than the noise level. Let $S_0 = \{(i,j) \in \{1,\dots,p\}^2 : \Theta^0_{ij} \neq 0\}$ be the non-zero set of Θ_0, let $S^0_\tau = \{(i,j) \in \{1,\dots,p\}^2 : |\Theta^0_{ij}| > \tau_{ij}\}$ and $\hat{S}_\tau = \{(i,j) \in \{1,\dots,p\}^2 : |\hat{T}_{ij}| > \tau_{ij}\}$ for some $\tau = (\tau_{ij})_{i,j=1,\dots,p}$.

Theorem 3. *Suppose that $X_1,\dots,X_n$ are i.i.d. $\mathcal{N}(0,\Sigma_0)$, where $\Theta_0 := \Sigma_0^{-1}$ exists, $\Theta_0 \in \mathcal{G}_n(s_n, L)$ where $L = \mathcal{O}(1)$ and $s_n = o(\sqrt{n}/\log p)$. Let $\hat{T}$ be the estimator defined in (2), let $\lambda_j \asymp \sqrt{\log p/n}$ uniformly in $j = 1,\dots,p$, $\tau_{ij} \asymp \sigma_{ij}\sqrt{\log p/n}$ and $\hat{\tau}_{ij} \asymp \hat{\sigma}_{ij}\sqrt{\log p/n}$, $i,j = 1,\dots,p$ are suitably chosen. Then*

$$\lim_{n\to\infty} \mathbb{P}(S^0_\tau \subset \hat{S}_{\hat{\tau}} \subset S_0) = 1.$$

Theorem 3 implies that by thresholding $\hat{T}$, all variables which are sufficiently high above the noise level are selected and guarantees no false positives.
We remark that the normality assumption in Theorems 1, 2 and 3 may be relaxed to (uniform) sub-Gaussianity of the margins of the underlying distribution. More precisely, it suffices to assume the following.

Condition 1. *Let $X_1,\dots,X_n \in \mathbb{R}^p$ be independent, $\mathbb{E}X_j = 0, \mathrm{Cov}(X_j) = \Sigma_0$ for $j = 1,\dots,n$. Suppose that for some constant $K = \mathcal{O}(1)$ it holds*

$$\max_{j=1,\dots,n} \sup_{\alpha\in\mathbb{R}^p:\|\alpha\|_2\leq 1} \mathbb{E}e^{|\alpha^T X_j|^2/K^2} = \mathcal{O}(1).$$

The results of Theorems 1, 2 and 3 may be regenerated when the Gaussianity assumption is replaced by Condition 1. Note that in Theorem 1, one needs to replace $\hat{\sigma}^2_{ij}$ by a consistent estimator of the asymptotic variance $\sigma^2_{ij} = \mathrm{Var}((\Theta^0_i)^T X_1 X_1^T \Theta^0_j)$, where Θ^0_i is the i-th column of Θ_0.
Finally, note that the proposed approach for de-biasing a penalized estimator is not limited to the estimator in [3]. For instance, analogous approach may be applied to obtain an asymptotically normal estimator and confidence regions based on the graphical Lasso, which is treated in detail in [2].

References

[1] Bickel, P. J., Klaassen, J., Wellner, J. A., and Ritov, Y. *Efficient and adaptive estimation for semiparametric models*. Springer (1993).

[2] Jankova, J. and van de Geer, S. Confidence intervals for high-dimensional inverse covariance estimation. *ArXiv :1403.6752* (2014).

[3] Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**(3) (2006), 1436–1462.

[4] Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. Asymptotic normality and optimalities in estimation of large Gaussian graphical model. *ArXiv: 1309.6024* (2013).

[5] van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, **42**(3) (2014), 1166-1202.

[6] Zhang, C.-H. and Zhang, S. S. Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society: Series B*, **76** (2014), 217–242.

Quantum versions of the randomization criterion

Anna Jenčová

A classical statistical experiment is a parametrized family $\mathcal{E} = (X, \Sigma, \{p_\theta,\ \theta \in \Theta\})$ of probability densities on a sample space X. Let $\mathcal{F} = (Y, \Sigma, \{q_\theta,\ \theta \in \Theta\})$ be another experiment, we assume that X, Y and Θ are finite sets. The classical randomization criterion due to Blackwell [1] and Törgersen [9], states that ϵ-deficiency of $\mathcal{E}$ with respect $\mathcal{F}$ in the sense of [1] is equivalent to

$$\inf_M \sup_{\theta\in\Theta} \|M(p_\theta) - q_\theta\|_1 \leq 2\epsilon, \tag{1}$$

where the infimum is taken over all Markov kernels $M : X \times Y \to [0, 1]$. Here ϵ-deficiency is defined by comparison of the average risks of decision rules for the two experiments, for all decision problems.

A quantum statistical experiment is a family $\mathcal{E} = (\mathcal{H}, \{\rho_\theta, \theta \in \Theta\})$ of density operators on a Hilbert space $\mathcal{H}$, we assume that $|\Theta| = n$ and $\dim(\mathcal{H}) < \infty$. Let $\mathcal{F} = (\mathcal{K}, \{\sigma_\theta, \theta \in \Theta\})$ be another experiment. In analogy with the classical case, we want to relate the quantity

$$\inf_{\alpha\in\mathcal{C}(\mathcal{H},\mathcal{K})} \sup_\theta \|\alpha(\rho_\theta) - \sigma_\theta\|_1 \tag{2}$$

to risks of decision rules for the two experiments, here $\mathcal{C}(\mathcal{H}, \mathcal{K})$ is the set of completely positive trace preserving maps, or *channels*, $B(\mathcal{H}) \to B(\mathcal{K})$, representing physically meaningful transformations of quantum states. Besides the classical decision problems, we may consider the quantum ones, introduced by Matsumoto [7]. These are triples $(\mathcal{D}, W, \mathcal{E})$, where $\mathcal{D}$ is a Hilbert space ($\dim(\mathcal{D}) < \infty$) and $W : \theta \mapsto W_\theta \in B(\mathcal{D})^+$. It is clear that classical decision spaces correspond precisely to the case when W_θ are mutually commuting.

More generally, we may consider a pair of channels with the same input space, $\Phi \in \mathcal{C}(\mathcal{H}_0, \mathcal{H})$ and $\Psi \in \mathcal{C}(\mathcal{H}_0, \mathcal{K})$, and the quantity

$$\inf_{\alpha\in\mathcal{C}(\mathcal{H},\mathcal{K})} \|\alpha \circ \Phi - \Psi\|_\diamond, \tag{3}$$

where $\|\cdot\|_\diamond$ is the diamond norm introduced in [6] as a distinguishability norm for channels. If Φ has the form $\Phi^{cq}_{\mathcal{E}} : B(\mathbb{C}^n) \ni A \mapsto \sum_\theta A_{\theta\theta}\rho_\theta \in B(\mathcal{H})$ and similarly $\Psi = \Phi^{cq}_{\mathcal{F}}$, then (3) is the same as (2). We next introduce the *post-processing decision problems* for quantum channels.

Let $\mathcal{L} = \mathcal{L}(\mathcal{H}_0, \mathcal{D})$ be the space of all hermitian linear maps $B(\mathcal{H}_0) \to B(\mathcal{D})$. With the cone CP of completely positive maps, $(\mathcal{L}, CP)$ becomes an ordered vector space. We identify its dual by $\mathcal{L}^* \equiv \mathcal{L}(\mathcal{D}, \mathcal{H}_0)$ with duality

$$\langle\phi, \psi\rangle = s(\psi \circ \phi) = \sum_{ij} \langle i, \psi \circ \phi(|i\rangle\langle j|)j\rangle$$

for some orthonormal basis $|i\rangle$ of $\mathcal{H}_0$. Then $CP^* = CP$, moreover, $\mathcal{C}(\mathcal{H}_0, \mathcal{D})$ is a *base section* in CP and $\|\cdot\|_\diamond$ is the corresponding norm, [4, 5]. Let $\|\cdot\|^\diamond$ be the dual norm. A *post-processing decision space* is a pair $(\mathcal{D}, \Gamma)$, where $\Gamma \in CP(\mathcal{D}, \mathcal{H}_0)$. The decision space is called *classical* if Γ is a cq-map: $\Gamma = \Phi^{cq}_G$ for

some $G = (G_1, \ldots, G_{\dim(\mathcal{D})}) \subset B(\mathcal{H}_0)^+$. A decision rule for $(\mathcal{D}, \Gamma, \Phi)$ is a channel $\phi \in \mathcal{C}(\mathcal{H}, \mathcal{D})$ and its risk is given by $R_\Phi(\mathcal{D}, \Gamma, \phi) = \langle \phi \circ \Phi, \Gamma \rangle$.

Definition 1. *We say that Φ is ϵ-post-processing deficient with respect to Ψ, in notation $\Phi \succeq_\epsilon \Psi$, if for any post-processing decision space $(\mathcal{D}, \Gamma)$ and any $\psi \in \mathcal{C}(\mathcal{K}, \mathcal{D})$ there is some $\phi \in \mathcal{C}(\mathcal{H}, \mathcal{D})$ such that*

$$\langle \phi \circ \Phi, \Gamma \rangle \leq \langle \psi \circ \Psi, \Gamma \rangle + \epsilon \|\Gamma\|^\diamond$$

If this is holds restricted to classical $(\mathcal{D}, \Gamma)$, we say that Φ is classically ϵ-post-processing deficient with respect to Ψ, in notation $\Phi \succeq^c_\epsilon \Psi$.

The following general post-processing randomization criterion can be proved using the minimax theorem (as in the classical case) and duality of the two norms.

Theorem 1. *The following are equivalent.*

(i) $\Phi \succeq_\epsilon \Psi$
(ii) *For any* $\Gamma \in CP(\mathcal{K}, \mathcal{H})$, $\|\Psi \circ \Gamma\|^\diamond \leq \|\Phi \circ \Gamma\|^\diamond + \epsilon \|\Gamma\|^\diamond$
(iii) $\inf_{\alpha \in \mathcal{C}(\mathcal{H}, \mathcal{K})} \|\alpha \circ \Phi - \Psi\|_\diamond \leq 2\epsilon$

This result was proved in [5] in a more general setting, with the cone CP replaced by a cone $\mathcal{P}$ of positive maps, satisfying some natural properties. For example, $\mathcal{P}$ can be CP, the cone of all positive maps, or k-positive maps. For experiments, we obtain

Theorem 2. *The following are equivalent.*

(i) $\mathcal{E} \succeq_{\epsilon, \mathcal{P}} \mathcal{F}$
(ii) *For any* $\theta \mapsto W_\theta \in B(\mathcal{K})^+$, $\|\Phi_{W,\mathcal{F}}\|^\diamond_\mathcal{P} \leq \|\Phi_{W,\mathcal{E}}\|^\diamond_\mathcal{P} + \epsilon \|W\|$, *where* $\Phi_{W,\mathcal{E}}(A) = \sum_\theta \operatorname{Tr}(W_\theta A)\rho_\theta$, $\|W\| = \sum_\theta \|W_\theta\|$
(iii) $\inf_{\alpha \in \mathcal{C}_\mathcal{P}} \sup_\theta \|\alpha(\rho_\theta) - \sigma_\theta\|_1 \leq 2\epsilon$, *where $\mathcal{C}_\mathcal{P}$ is the set of trace preserving maps in $\mathcal{P}(\mathcal{H}, \mathcal{K})$.*

In particular, for $\mathcal{P} = CP$, this was obtained by Matsumoto [7]. Note that for any choice of $\mathcal{P}$, all operator-valued loss functions have to be considered in (ii), only the norm $\|\cdot\|^\diamond_\mathcal{P}$ depends on the cone. This suggests that classical deficiency is not sufficient for randomization criterion even if α is only required positive, as in fact was shown in [8]. In general, the relation of classical deficiency to (2) or (3) is not clear, but we have the following extension of a result of Buscemi [2].

Theorem 3. *For any $\epsilon \geq 0$, $\Phi \succeq_\epsilon \Psi$ if and only if $\Phi \otimes id_\mathcal{K} \succeq^c_\epsilon \Psi \otimes id_\mathcal{K}$. If $\xi \in \mathcal{C}(\mathcal{K}_0, \mathcal{K})$ is surjective, then $\Phi \succeq_0 \Psi$ if and only if $\Phi \otimes \xi \succeq^c_0 \Psi \otimes \xi$.*

Finally, we may consider pre-processings of channels in a similar way and obtain a corresponding randomization criterion. In particular, the results for POVM's, which are a special kind of channels, are related to *cleanness* defined in [3].

References

[1] D. Blackwell, *Comparison of experiments*, Proc. 2nd Berkeley Symp. on Math. Stat. and Probab. (1951), 93–102

[2] F. Buscemi, *Comparison of Quantum Statistical Models: Equivalent Conditions for Sufficiency*, Commun. Math. Phys. **310** (2012), 625–647
[3] F. Buscemi et al. *Clean positive operator valued measures*, J. Math. Phys. **46** (2005), 082109
[4] A. Jenčová, *Base norms and discrimination of generalized quantum channels*, J. Math. Phys. **55** (2014), 022201
[5] A. Jenčová, *Randomization theorems for quantum channels*, arXiv:1404.3900 (2014)
[6] A. Kitaev, *Quantum computations: Algorithms and error correction*, Russian Mathematical Surveys **52** (1997), 1191–1249
[7] K. Matsumoto, *A quantum version of randomization condition*, arXiv:1012.2650 (2010)
[8] K. Matsumoto, *An example of a quantum statistical model which cannot be mapped to a less informative one by any trace preserving positive map*, arXiv:1409.5658 (2014)
[9] E. Torgersen, *Comparison of statistical experiments when the parameter space is finite*, Z. Wahrscheintlichkeitstheorie verw. geb. **16** (1970), 219–249

When is an input state always better than the others?

Keiji Matsumoto

Statistical estimation and test of unknown channels have attracted interests of many researchers. Below, let $\{\Lambda_\theta\}_{\theta\in\Theta}$ be a family of unknown channels, where $\theta \in \Theta$ is the unknown parameter. In optimizing the process of inference, one has to optimize not only the measurement performed upon the output state $\Lambda_\theta \otimes \mathbf{I}(\rho_{in})$, but also the input state ρ_{in}.

In general, optimal input states depend on whether we are estimating state or testing hypothesis about unknown channels; they also depend on error measure, and detail of the setting (Bayesian, minimax, unbiased estimation, Neyman-Pearson test, etc.).

In some cases, however, the situation is less complicated. For example, [2] deals with estimation of group transform $\{U_g\}_{g\in\mathcal{G}}$, where $g \to U_g$ is a representation of the group $\mathcal{G}$ and g is unknown and to be estimated. They had shown that there is an input state which is optimal with respect to any $\mathcal{G}$-invariant loss functions. (In case of $\mathcal{G} = \mathrm{SU}(d)$ and $U_g = g$, maximally entangled states between the input space and the auxiliary space are optimal.) Meantime, [3] treats estimation of $\mathrm{SU}(2)$ channel by an unbiased estimator, and 'the loss function' here is the mean square error matrix of the estimate $\hat{\theta}$ of the unknown real vector θ which parameterizes $\mathcal{G} = \mathrm{SU}(2)$. Since the space of matrices is not totally ordered, the existence of the minimum is non-trivial. Put differently, if the loss is scalar valued increasing function of a mean square error matrix, then, maximally entangled states are optimal. Also, [8] studies discrimination of a pair of generalized Pauli matrices, and shows maximally entangled states minimize Bayesian error probability for any prior distributions. In case of qubits, they extended their result to minimax error probability [10]. Another example of such study is [11], where discrimination of two unitary operation is discussed. They found that minimizers of Bayesian error probability and the error probability of unambiguous discrimination are the same.

These results motivate the following definition: we say the input is *universally optimal* for the family $\{\Lambda_\theta\}_{\theta\in\Theta}$, roughly speaking, if it is optimal for all the statistical inferences and for all the loss functions. (The rigorous definition will be given

later.) We show that a universally optimal state exists (not necessarily uniquely) in case of group covariant and contravariant channels, unital qubit channels and some measurement families.

To prove these results, we have recourse to the theory of "comparison of state families" [1][7]; we write $\{\rho_\theta\}_{\theta\in\Theta} \succeq^c \{\sigma_\theta\}_{\theta\in\Theta}$ if the family $\{\rho_\theta\}_{\theta\in\Theta}$ is more informative than another family $\{\sigma_\theta\}_{\theta\in\Theta}$ with respect to any kind of statistical inferences. Then, our target is to prove the existence of the state $\rho_{\rm opt}$ with

$$\forall\rho' \ \{(\Lambda_\theta\otimes\mathbf{I})(\rho_{\rm opt})\}_{\theta\in\Theta} \succeq_c \{(\Lambda_\theta\otimes\mathbf{I})(\rho')\}_{\theta\in\Theta}$$

Based on these results, some related topics are discussed. The first topic is effect of entanglement between the input space and the auxiliary space. For example, in [8][9][10], they study the condition that Bayes risk and minimax risk of discrimination of two unital qubit channels is smaller on an entangled state than on any separable state. In our case, it is shown that a maximally entangle is universally optimal for some channel families. But there might be a separable state which is as good as maximally entangled states. So we question whether the entanglement is really needed or not. The second topic discussed is the existence of universally optimal states under the setting where the given channel can be used for several times.

References

[1] F. Buscemi, "Comparison of Quantum Statistical Models: Equivalent Conditions for Sufficiency", Communications in Mathematical Physics Vol. 310, No. 3, 625-647 (2012)

[2] G. Chiribella, G. M. D'Ariano, and M. F. Sacchi, Optimal estimation of group transformations using entanglement, Phys. Rev. A 72 042338 (2005)

[3] A. Fujiwara, "Estimation of SU(2) operation and dense coding: An information geometric approach," Phys. Rev. A, vol. 65, 012316 (2002)

[4] A. Fujiwara and Hiroshi Imai, "Quantum parameter estimation of a generalized Pauli channel," J. Phys. A: Math. Gen., vol. 36, pp. 8093-8103 (2003)

[5] A. Fujiwara, "Estimation of a generalized amplitude-damping channel," Phys. Rev. A, vol. 70, 012317 (2004)

[6] A. Fujiwara and H. Imai, "A fibre bundle over manifolds of quantum channels and its application to quantum statistics," J. Phys. A: Math. Theor., vol. 41, 255304 (2008)

[7] K. Matsumoto, "A quantum version of randomization criteria" (2010)

[8] M. F. Sacchi, "Optimal discrimination of quantum operations", Phys. Rev. A 71, 062340 (2005)

[9] M. F. Sacchi, "Minimum error discrimination of Pauli channels", J. Opt. B 7, S333 (2005).

[10] M. F. Sacchi, "Entanglement can enhance the distinguishability of entanglement-breaking channels", Phys. Rev. A 72, 014305 (2005)

[11] M. Ziman, "Single-shot discrimination of quantum unitary processes", Journal of Modern Optics, Vol. 57, No. 3, pp. 253-259 (2010)

Optimal classification and nonparametric regression for functional data

Alexander Meister

We establish minimax convergence rates for classification of functional data and for nonparametric regression with functional design variables. The optimal rates are of logarithmic type under smoothness constraints on the functional density and the regression mapping, respectively. These asymptotic properties are attainable by conventional kernel procedures. The bandwidth selector is automatically adaptive. In this work the functional data are considered as realisations of random variables which take their values in a general Polish metric space. We impose certain metric entropy constraints on this space; but no algebraic properties are required.

References

[1] A. Meister, *Optimal classification and nonparametric regression for functional data*, under review.

Experimental, encoded quantum computation: statistical and mathematical challenges, right now

Thomas Monz

Advances towards fault-tolerant quantum computation currently suggest that quantum information ought to be encoded in and distributed over several physical qubits. Here, the encoding should be efficient, enable error correction, and allow for manipulation of the quantum information directly within the code-space. One particular promising approach is to topologically encode quantum information. Here, quantum information is encoded such that only global properties of the system encode information. A topological encoding thus, intrinsically, renders the system resistant to local perturbations. In the following we will focus on a particular encoding: the topological colour code [1, 2]. One prominent feature of the colour code is that the entire Clifford gate set can be implemented in a transversal way. Here, transversality means that any Clifford gate acting on the logical, encoded qubit, can be implemented by applying the very same operation locally on the substituting individual, physical qubits. This property facilitates the implementation of logical gates in a physical realisation. In addition, the aspect that gates only need to be applied on the individual substituting qubits, notably affects error propagation properties and results in high error thresholds on the order of 1%. An additional useful feature is that the colour code belongs to the class of Calderbank-Shor-Stean (CSS) stabiliser codes. Here, X/Z errors can be detected independently, and manifest themselves in the according Z/X stabilisers of the code. In total, any first realisation of a logical qubit should thus demonstrate the key features of the encoding: the application of gates directly on the encoded qubit, and the successful detection of errors.

For the first physical realisation of a topological qubit we focused on the smallest logical instance consisting of 7 physical qubits, implemented in a system of

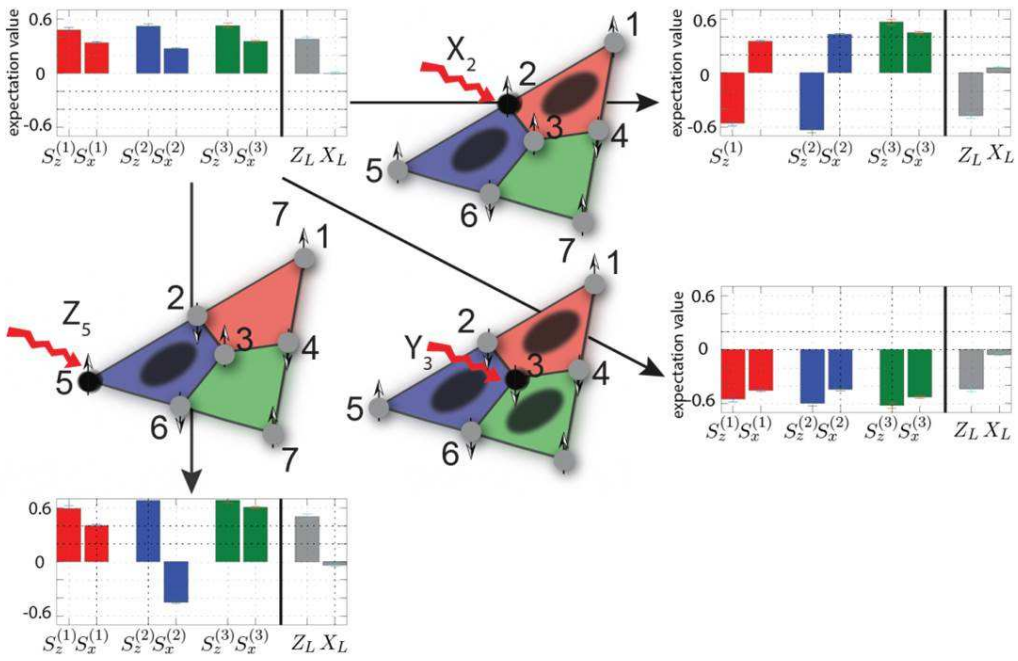

FIGURE 1. Topological colour code subject to errors: The effect of X/Z errors on single qubits clearly the corresponding stabilisers in Z/X. Based on the information of the negative stabilisers, the location and property of the error can be obtained. Thus, any error can be uniquely identified, and later on corrected. The figure above shows how an X error on the second qubit, that has overlap both with the red and blue plaquette, affects the Z stabiliser values of said plaquettes. In a similar fashion, Z and Y errors are demonstrated and identified on qubits 5 and 3.

trapped Ca-ions [3]. Given the CSS nature of the encoding, the state is an eigenstate of 3 Z and X stabilisers, acting on trivalent coloured plaquettes (see Fig. 1). In our realisation we step-wise initialise the logical state plaquette after plaquette. Subsequently, we demonstrate the error detection capabilities of the code by looking at the expectation values of the different X- and Z-stabiliser of the individual plaquettes. The effect of an error, namely the change in the sign of the expectation value of the affected plaquettes, is clearly demonstrated in Fig. 1.

In addition we have been able to apply all Clifford gate operations on the logical qubit, investigate the coherence properties of the logical qubit and perform preliminary evaluations on leakage from the code space affecting the qubit during logical operations. All details about the presented work can be found in Ref. [4]

This work, however, should be considered to motivate discussions in the interdisciplinary field of quantum theory, mathematics, and statistics. The presented system, though only consisting of 7 qubits, exceeds the majority of characterisation methods available to date: the largest quantum state ever subject to full tomography consisted of 8 qubits, process tomography has only been performed on

up to 3 qubits. It follows that any attempt to perform process tomography on the 7-qubit operations exceeds the current capabilities by orders of magnitude. While one might think about applying model-specific tomography methods (for instance to local, non-entangling operations only), such models would require independent validation tests to be applied onto the experimental data. Alternatively one could make use of "Randomized Benchmarking" (RB) [5] on the logical qubit, as the entire Clifford gate set is directly available in the experiment.

But also this approach has road-blocks: So far, RB has only been considered in leakage-free, Markovian systems. In the presented experiment it is not clear to which degree the system is Markovian (as there exist hardly any experimentally applicable measure for it). Another aspect is that, so far, Markovianity (or rather correlations) have many been considered in time. For encoded qubits there may also be spatial correlations, which are, for instance, employed in the field of decoherence-free subspaces. It is not clear how such spatial correlations may affect the validity of numbers derived from RB. Finally, there is also leakage from the code-space to the Hilbert-space of the 7 qubits. This effect has thus far not been considered the field of RB. Here, leakage needs to be separately taken into account with respect to other errors resulting from the applied gates. Alternatively to RB one could try to look into methods such as presented by Robin Blume-Kohout at the workshop, but also these need to be extended with respect to temporal and spatial correlations as well as leakage.

The presented work can safely conclude that experimental progress is currently held back by both numerical, analytical and statistical tools to truthfully evaluate the system-performance of already moderately-sized systems such as the presented 7-qubit experiment, even more so in the context of logical systems [4]. Within the framework of the workshop, several hours have been spend on discussing potential solutions for the presented problems - some of them looking promising and within current capabilities of experimentalists.

References

[1] H. Bombin, M. A. Martin-Delgado, *Topological quantum distillation*, Phys. Rev. Lett. **97**, 180501 (2006)

[2] H. Bombin, M. A. Martin-Delgado, *Topological computation without braiding*, Phys. Rev. Lett. **98**, 160502 (2007).

[3] P. Schindler, et al., *A quantum information processor with trapped ions*, New. J. Phys. **15**, 123012 (2013)

[4] D. Nigg, et al., *Quantum computations on a topologically encoded qubit*, Science **345**, 302–305 (2014)

[5] J. Emerson, et al., *Symmetrized Characterization of Noisy Quantum Processes*, Science **317**, 1893-1896 (2007)

Nonparametric regression with one-sided error distribution

Natalie Neumeyer
(joint work with Holger Drees and Leonie Selk)

We consider a nonparametric boundary regression model

$$Y_i = g(\tfrac{i}{n}) + \varepsilon_i, \quad i = 1, \dots, n,$$

where the (unobserved) errors $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed with cumulative distribution function F that satisfies

$$F(y) = 1 - c|y|^\alpha + o(|y|^\alpha)$$

for some $\alpha > 0$ for $y \nearrow 0$. The regression function g is assumed to belong to some Hölder class of order $\beta \in (0, \infty)$. We estimate the regression function via minimization of the local integral of a polynomial approximation. More specifically, for $x \in [0,1]$ let $\hat{g}(x) = p(x)$, where p is a polynomial of order $\lceil \beta \rceil - 1$ and minimizes

$$\int_{x-h_n}^{x+h_n} p(t)\, dt$$

under the constraints $p(\frac{j}{n}) \geq Y_j$ for all $j \in \{1, \dots, n\}$ such that $|\frac{j}{n} - x| \leq h_n$. Here $(h_n)_n$ is a sequence of bandwidths with $h_n \to 0$, $nh_n/|\log h_n| \to \infty$. We obtain the following uniform rate of convergence,

$$\sup_{x \in [h_n, 1-h_n]} |\hat{g}(x) - g(x)| = O(h_n^\beta) + O_P\Big(\Big(\frac{|\log h_n|}{nh_n}\Big)^{1/\alpha}\Big).$$

The minimal rate $O_P((\log n/n)^{\beta/(\alpha\beta+1)})$ obtained from $h_n \sim ((\log n)/n)^{1/(\alpha\beta+1)}$ is faster than the typical rate $O_P((\log n/n)^{\beta/(2\beta+1)})$ in nonparametric mean regression with regular error distribution if the error distribution is irregular in the sense that sufficient mass is concentrated near the endpoint (i. e. $\alpha \in (0,2)$).
For inference on the error distribution let $\hat{F}_n$ denote the empirical distribution function of residuals $\hat{\varepsilon}_i = Y_i - \hat{g}(\frac{i}{n})$, $i = 1, \dots, n$. If the error distribution F is Hölder continuous of order $\alpha \wedge 1$ and $\beta^{-1} < \alpha < 2 - \beta^{-1}$ holds, the influence of the regression estimation is negligible, i. e. $n^{1/2}(\hat{F}_n - F)$ converges weakly to a Brownian bridge composed with F. This result is remarkably different from corresponding results on the residual-based empirical distribution function in mean regression models. It is true for all cases of irregularity $\alpha \in (0,2)$ if the regression function is smooth enough, and it can readily be used for goodness-of-fit testing. Applying a bias-reduced and smoothed version of the regression estimator $\hat{g}$ we can even extend the result (under suitable choice of smoothing parameters) to the case $\alpha < 3 - 3/(2\beta)$ when we either stay away from the boundary or assume a bounded error density. Details can be found in [1].

References

[1] H. Drees, N. Neumeyer, L. Selk *Hypotheses tests in boundary regression models*, arXiv:1408.3979 (2014).

Uncertainty quantification and confidence sets in high-dimensional statistical models

Richard Nickl
(joint work with Sara van de Geer)

The problem of constructing confidence sets in the high dimensional linear model with n response variables and p parameters, possibly $p \geq n$, is considered. Full honest adaptive inference is possible if the rate of sparse estimation does not exceed $n^{-1/4}$, otherwise sparse adaptive confidence sets exist only over strict subsets of the parameter spaces for which sparse estimators exist. Necessary and sufficient conditions for the existence of confidence sets that adapt to a fixed sparsity level of the parameter vector are given in terms of minimal ℓ^2-separation conditions on the parameter space. The design conditions cover common coherence assumptions used in models for sparsity, including (possibly correlated) sub-Gaussian designs. The proof techniques are inspired by Hoffmann and Nickl (2011).

References

[1] R. Nickl, S. van de Geer, *Confidence sets in sparse regression*, Annals of Statistics **41** (2013), 2852-2876.
[2] M. Hoffmann, R. Nickl, *On adaptive inference and confidence bands*, Annals of Statistics **39** (2011), 2383-2409.

Optimal error intervals for quantum parameter estimation

Jiangwei Shang
(joint work with Xikun Li, Hui Khoon Ng, Berthold-Georg Englert)

Information-theoretic quantities like purity, fidelity, entanglement, *etc.* have played important roles in our understanding of quantum information, quantum communication and quantum computation. However, these quantities are mathematical constructs, with no associated ability to measure them directly in the laboratory. The conventional wisdom is to perform full state reconstruction for the quantum system and then compute the quantities of interest from the estimated state. Full state reconstruction is feasible for small quantum systems, but rapidly defies the best analytical and numerical efforts as the dimensionality of the system grows. It thus becomes desirable to design new schemes to *directly* estimate various quantities of interest, without first going through full state reconstruction.

In our previous work, we have found a very simple construction of optimal error regions for quantum state estimation [1]. A point estimator, constructed from the measurement outcomes on a finite set of independently and identically prepared systems, can never be perfectly accurate; it has to be supplemented with an error region that summarizes our uncertainty about the guess. Exploiting the natural correspondence between the size of a region in state space and its prior content, we showed that the optimal choices for two types of error regions—the maximum-likelihood region, and the smallest credible region—are both concisely described

as the set of all states for which the likelihood (for the given tomographic data) exceeds a threshold value, *i.e.*, a bounded-likelihood region. These error regions are reminiscent of the standard error regions obtained by analyzing the vicinity of the maximum of the likelihood function, a construction valid only when a large number of copies of the state have been observed. Yet, we require no such restriction. Our error regions are conceptually different from confidence regions, a subject of recent discussion in the context of quantum state estimation; however, they can serve as good starting points for constructing confidence regions.

Now, we are interested in extending our construction for error regions to direct estimation of parameters of interest [2]. Moreover, we know that the best error bars one can write down for estimating parameters from the tomography measurements do not usually come from the error regions one first constructs for the estimator for the state; see Fig. 1 for an example. The immediate task is then to extend our methods for state estimation to parameter estimation, which becomes possible by employing the numerical tools that we have developed for sampling in the quantum state space [3, 4].

In Ref. [2], we propose a systematic method to construct optimal error intervals for quantum parameter estimation. For given data D, we look for the smallest credible interval (SCI) with a pre-chosen value of credibility. We show that the SCI contains all parameters with the likelihood conditional on the parameter exceeding a certain threshold, which is specified by a fraction of the maximum value of the likelihood, *i.e.*, a bounded-likelihood interval. Surprisingly, we find that the results obtained for parameter estimation take very similar forms as those for state estimation. Specifically, we construct SCIs for the purity and fidelity (with respect to certain target states) of single-qubit states, as well as for the CHSH quantity of two-qubit states.

References

[1] J. Shang, H. K. Ng, A. Sehrawat, X. Li, and B.-G. Englert, *Optimal error regions for quantum state estimation*, New J. Phys. **15** (2013), 123026.

[2] X. Li, J. Shang, H. K. Ng, and B.-G. Englert, *Optimal error intervals for quantum parameter estimation*, in preparation (2014).

[3] J. Shang, Y.-L. Seah, H. K. Ng, D. J. Nott, and B.-G. Englert, *Monte Carlo integration over regions in the quantum state space. I*, arXiv:1407.7805 [quant-ph] (2014).

[4] Y.-L. Seah, J. Shang, H. K. Ng, D. J. Nott, and B.-G. Englert, *Monte Carlo integration over regions in the quantum state space. II*, arXiv:1407.7806 [quant-ph] (2014).

Multiplier bootstrap for confidence estimation

Vladimir Spokoiny
(joint work with Mayya Zhilova)

In this talk we consider a multiplier bootstrap procedure for the problem of confidence estimation using a quasi maximum likelihood method. A confidence set is based on a likelihood function taken for a rather general parametric model. A radius of the confidence set is determined by a multiplier bootstrap. The aim of

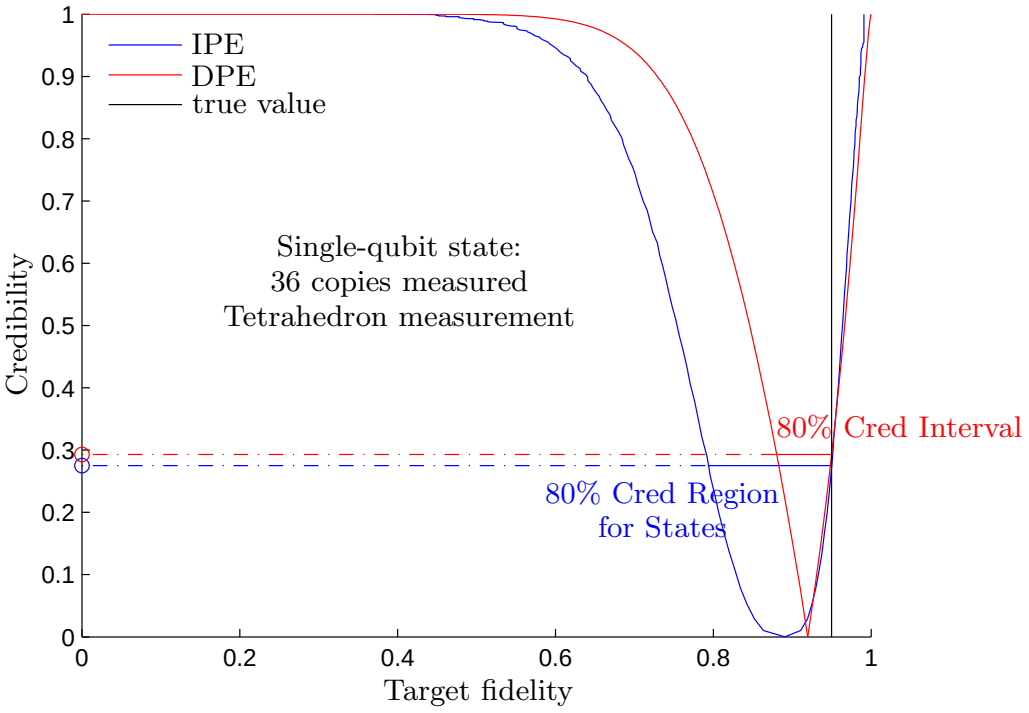

FIGURE 1. Optimal error intervals for fidelity (with respect to certain target states) of single-qubit states by direct parameter estimation (DPE, red curve) and indirect parameter estimation (IPE, blue curve) respectively. The true fidelity of 0.95 is indicated by the vertical black line. The dotted-dashed horizontal lines mark out the smallest intervals with 80% credibility that contain the true fidelity.

the study is to check the validity of the bootstrap procedure in situations with a large parameter dimension, a limited sample size and a possible misspecification of the parametric assumption.

Let the data sample $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ consist of independent random observations and belong to the probability space $(\Omega, \mathcal{F}, I\!P)$. We do not assume that the observations Y_i are identically distributed, moreover, no specific parametric structure of $I\!P$ is being required. Consider some known regular parametric family $\{I\!P_\theta\} \stackrel{\text{def}}{=} \{I\!P_\theta \ll \mu_0, \theta \in \Theta \subset I\!R^p\}$ with parameter's dimension p. This family induces the log-likelihood process $L(\theta)$ of the sample $\mathbf{Y}$:

$$L(\theta) \stackrel{\text{def}}{=} \log \left\{ \frac{dI\!P_\theta}{d\mu_0}(\mathbf{Y}) \right\}.$$

The (quasi) maximum likelihood estimator $\tilde{\theta}$ and the target parameter θ^* are defined as follows:

$$\tilde{\theta} \stackrel{\text{def}}{=} \operatorname{argmax}_{\theta \in \Theta} L(\theta), \qquad \theta^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\theta \in \Theta} I\!E L(\theta).$$

Define the likelihood-based confidence set as:

$$\mathcal{E}(\mathfrak{z}) \stackrel{\text{def}}{=} \{\theta : L(\tilde{\theta}) - L(\theta) \le \mathfrak{z}^2/2\}.$$

Here the parameter $\mathfrak{z} \ge 0$ determines the size of the confidence set. Let $1-\alpha \in (0,1)$ be a fixed required confidence level. We are interested in finding a minimal possible value of $\mathfrak{z}$ such that it holds:

$$I\!P\{\theta^* \in \mathcal{E}(\mathfrak{z})\} \ge 1-\alpha,$$

this is equivalent to the estimation of the upper α-quantile of the likelihood ratio statistic $L(\tilde{\theta}) - L(\theta^*)$. For this purpose we consider multiplier bootstrap (or weighted bootstrap) procedure. The idea of the procedure is to mimic a distribution of the likelihood ratio statistic by reweighing its summands with random multipliers independent of the data. Let $u_1, \ldots, u_n$ be scalar i.i.d. random variables independent of $\mathbf{Y}$ with continuous c.d.f., $I\!E u_1 = 1$, $\operatorname{Var} u_1 = 1$, $I\!E \exp(u_1) < \infty$ (e.g. $\mathcal{N}(1,1)$, $exp(1)$). Multiply the summands of the likelihood function $L(\theta)$ with the new random variables:

$$L^{\circ}(\theta) \stackrel{\text{def}}{=} \sum\nolimits_{i=1}^{n} \log\left\{\frac{dI\!P_\theta^{\circ}}{d\mu_0}(Y_i)\right\} u_i.$$

Here the probability distribution is taken conditionally on the data $\mathbf{Y}$, which is denoted by the sign $^{\circ}$. The multiplier bootstrap induces the probability space conditional on the data $\mathbf{Y}$. A simple but important observation is that $I\!E^{\circ} L^{\circ}(\theta) \equiv I\!E\big[L^{\circ}(\theta) \,\big|\, \mathbf{Y}\big] = L(\theta)$, and hence,

$$\operatorname{argmax}_{\theta\in\Theta} I\!E^{\circ} L^{\circ}(\theta) = \operatorname{argmax}_{\theta\in\Theta} L(\theta) = \tilde{\theta}.$$

In other words, the target parameter in the bootstrap world is precisely known and it coincides with the maximum likelihood estimator $\tilde{\theta}$ conditioned on $\mathbf{Y}$, therefore, the bootstrap likelihood ratio statistic $L^{\circ}(\tilde{\theta}^{\circ}) - L^{\circ}(\tilde{\theta}) \stackrel{\text{def}}{=} \sup_{\theta\in\Theta} L^{\circ}(\theta) - L^{\circ}(\tilde{\theta})$ is fully computable and leads to a simple computational procedure for the approximation of the distribution of $L(\tilde{\theta}) - L(\theta^*)$.

The main results are given in Theorems 1, 2 below. The first statement requires a so called "small modeling bias" condition (SmB), formulated through a bound on a relation between the covariance matrices: $\operatorname{Var}\{\nabla_\theta L(\theta^*)\}$ and $\operatorname{Var}\{\nabla_\theta L^{\circ}(\theta^*) \,|\, \mathbf{Y}\}$. If the parametric family $\{I\!P_\theta\}$ is correct, then the modelling bias is equal to zero. (SmB) assumes that the true model does not deviate significantly from the parametric family. In this case the multiplier bootstrap procedure work if the relation p^3/n is small. If (SmB) does not hold (i.e. the deviation between the true model and $\{I\!P_\theta\}$ is large), then the multiplier bootstrap continues to apply but becomes a bit conservative: the size of the constructed confidence sets is increased by the modeling bias (see Theorem 2). The precise formulations of the results are given in [1]. They are illustrated with numerical experiments further in the text.

Theorem 1 (Validity of the bootstrap under a small modeling bias)**.**

(1) *It holds for all* $\mathfrak{z} \ge p$ *with probability* $\ge 1 - C_1 \mathrm{e}^{-\mathtt{x}}$, $C_1 \ge 12$, $\mathtt{x} > 0$

$$\left| I\!P\{L(\tilde{\theta}) - L(\theta^*) \ge \mathfrak{z}\} - I\!P^{\circ}\{L^{\circ}(\tilde{\theta}^{\circ}) - L^{\circ}(\tilde{\theta}) \ge \mathfrak{z}\} \right| \le \Delta$$

for a deterministic $\Delta \le C_2\{(p+\mathtt{x})^3/n\}^{1/8}$ *and a generic constant* C_2.

(2) *Moreover, for* $\mathfrak{z}_\alpha^\circ \stackrel{\text{def}}{=} \min\Big\{\mathfrak{z} \ge 0 \colon I\!P^\circ\Big(L^\circ(\tilde\theta^\circ) - L^\circ(\tilde\theta) \ge \mathfrak{z}\Big) = \alpha\Big\}$

$$\big|I\!P\{\theta^* \in \mathcal{E}(\mathfrak{z}_\alpha^\circ)\} - (1-\alpha)\big| \le \Delta.$$

Theorem 2 (Conservativeness of the bootstrap for a large modeling bias)**.** *It holds for all* $\mathfrak{z} \ge p$ *with probability* $\ge 1 - C_1 \mathrm{e}^{-\mathtt{x}}$

$$I\!P\big\{L(\tilde\theta) - L(\theta^*) \ge \mathfrak{z}\big\} - I\!P^\circ\big\{L^\circ(\tilde\theta^\circ) - L^\circ(\tilde\theta) \ge \mathfrak{z}\big\} \le \Delta$$

for Δ, C_1 *from Theorem 1.*

The proofs are based on non-asymptotic square-root Wilks expansions for the likelihood ratio statistics for both $\mathbf{Y}$ and bootstrap cases, non-asymptotic Gaussian approximation of the Euclidean norm of a random vector, and comparison of distributions of Gaussian vectors.

Numerical results. Below we show the results of numerical experiments illustrating the nice performance of the multiplier bootstrap procedure. In all the experiments we took the same number of samples: 10^4 data samples for estimation of quantiles of the likelihood ratio, 10^4 $\{u_i\}$ samples and 10^4 data samples for estimation of quantiles of the bootstrap likelihood ratio. The sample size is $n = 50$.

1) The first experiment checks, how well the procedure works in the case of the correct model. Let the data be i.i.d., $Y_i \sim \mathcal{N}(2,1)$, then $L(\theta) = -\sum_{i=1}^n (Y_i - \theta)^2/2$. Table 1 shows the effective coverage probabilities of the quantiles estimated using multiplier bootstrap. The second row contains the range of the nominal confidence levels: $0.99, \ldots, 0.75$. The first left column describes the distribution of the bootstrap weights: $\mathcal{N}(1,1)$ or $exp(1)$. The rows below the second one show the frequency of the event: "quantile of the real likelihood ratio $\le$ quantile of the bootstrap likelihood ratio".

TABLE 1. Coverage probabilities for the correct i.i.d. model

	Confidence levels					
$\mathcal{L}(u_i)$	**0.99**	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**
$exp(1)$	0.99	0.94	0.89	0.83	0.78	0.73
$\mathcal{N}(1,1)$	0.99	0.95	0.89	0.84	0.80	0.75

2) In the second experiment we consider constant regression model with misspecified heteroscedastic errors: $Y_i = 2 + \sigma_i \varepsilon_i$, $i = 1, \ldots, n$. The i.i.d. errors $\varepsilon_i \sim Lap(0, 2^{-1/2})$ s.t. $I\!E(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = 1$. The coefficients σ_i are deterministic: $\sigma_i \stackrel{\text{def}}{=} 0.5\,\{4 - i \pmod 4\}$. The quasi-likelihood function is the same as in the first experiment: $L(\theta) = -\sum_{i=1}^n (Y_i - \theta)^2/2$, i.e. it is misspecified, since it corresponds to the i.i.d. standard normal distribution. Table 2 describes the 2-nd experiment's results similarly to the table 1.

TABLE 2. Coverage probabilities for the misspecified heteroscedastic noise

	Confidence levels					
$\mathcal{L}(u_i)$	**0.99**	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**
$exp(1)$	0.98	0.93	0.87	0.82	0.77	0.72
$\mathcal{N}(1,1)$	0.98	0.94	0.88	0.83	0.78	0.73

3) In the third experiment we consider biased regression with misspecified i.i.d. errors: $Y_i = \beta \sin(X_i) + \varepsilon_i$, $\varepsilon_i \sim Lap(0, 2^{-1/2})$ i.i.d. The design points X_i are equidistant on $[0, 2\pi]$. Taking the likelihood function $L(\theta) = -\sum_{i=1}^{n}(Y_i - \theta)^2/2$ yields $\theta^* = 0$. Therefore, the larger is the amplitude $\beta > 0$, the bigger is bias of the mean constant regression. We consider two cases: $\beta = 0.25$ with fulfilled (SmB) condition and $\beta = 1.25$, when (SmB) does not hold.

TABLE 3. Coverage probabilities for the misspecified biased regression

		Confidence levels					
$\mathcal{L}(u_i)$	β	**0.99**	**0.95**	**0.90**	**0.85**	**0.80**	**0.75**
$\mathcal{N}(1,1)$	0.25	0.98	0.94	0.89	0.84	0.79	0.74
	1.25	1.0	0.99	0.97	0.94	0.91	0.87

REFERENCES

[1] V. Spokoiny, M. Zhilova, *Bootstrap confidence sets under a model misspecification*, arxiv: 1410.0347 (2014).

Higher order isotropy and lower bounds for sparse quadratic forms

SARA VAN DE GEER
(joint work with Alan Muro)

Let X be an $n \times p$ matrix with rows being i.i.d. copies of a random row-vector $X_0 \in \mathbf{R}^p$. The theoretical inner-product matrix is $\Sigma_0 := \mathbf{E}X_0^T X_0$ (assumed to exist). We let $\hat{\Sigma} := X^T X/n$ be the empirical inner-product matrix.

Let $m \geq 2$. The random vector X_0 is weakly m-th order isotropic with constant C if for all $u \in \mathbf{R}^p$ with $u^T \Sigma_0 u = 1$ it holds that

$$P(|X_0 u| > t) \leq (C/t)^m \ \forall \ t > 0.$$

We consider lower bounds for quadratic forms of the form $u^T \hat{\Sigma} u$, where $u \in \mathbf{R}^p$, $u^T \Sigma_0 u = 1$ and $\|u\|_1 \leq M$ with M some constant. We discuss extensions of [2] to a case with $p > n$ and of [3] to isotropy of any order $m > 2$. The isotropy conditions allow for refinements of results in [1] for compatibility constants and restricted eigenvalues.

A random variable Z is called Bernstein with constants σ and K if for all $k \in \{2, 3, \ldots\}$

$$\mathbf{E}|Z|^k \leq \frac{k!}{2} K^{k-2} \sigma^2.$$

Theorem 1 [4]. *Suppose the entries in X_0 are Bernstein with constants σ_X and K_X and that for some $m > 2$ the random vector X_0 is weakly m-th order isotropic with constant C_m. Define*

$$D_m := [2C_m]^{\frac{m}{m-1}}(m-1)/(m-2). \tag{1}$$

For all $t > 0$, with probability at least $1 - \exp[-t]$

$$\inf_{u^T \Sigma_0 u = 1,\ \|u\|_1 \leq M} u^T \hat{\Sigma} u \geq 1 - \Delta_n^L(M, t)$$

where

$$\Delta_n^L(M,t) := D_m \left(16 \min\left\{ M\delta_n, \sqrt{\frac{p}{n}} \right\} + \sqrt{\frac{2t}{n}} \right)^{\frac{m-2}{m-1}} + \frac{8 D_m^2}{3} \left(\frac{t}{n} \right)^{\frac{m-2}{m-1}} \tag{2}$$

with

$$\delta_n := \sigma_X \sqrt{\frac{2\log(2p)}{n}} + K_X \frac{\log(2p)}{n}.$$

We now denote for any subset $S \subset \{1, \ldots, p\}$ and for $u \in \mathbf{R}^p$, the vector

$$u_{j,S} := u_j 1\{j \in S\},\ j = 1, \ldots, p$$

and we let $u_{-S} := u - u_S$. We let for $s := |S|$ and $L > 0$ a constant

$$\phi_0^2(L, S) := \min\{ s u^T \Sigma_0 u :\ \|u_S\|_1 = 1,\ \|u_{-S}\|_1 \leq L \}$$

be the theoretical compatibility constant and

$$\hat{\phi}^2(L, S) := \min\{ s u^T \hat{\Sigma} u :\ \|u_S\|_1 = 1,\ \|u_{-S}\|_1 \leq L \}$$

be its empirical counterpart. These quantities play an important role in compressed sensing and oracle inequalities for ℓ_1-regularized estimators.

Lemma [4]. *Under the conditions of Theorem 1 and using its notation we find that for all $t > 0$, with probability at least $1 - \exp[-t]$*

$$\frac{\hat{\phi}^2(L,S)}{\phi_0^2(L,S)} \geq 1 - \Delta_n^L((L+1)\sqrt{s}/\phi_0(L,S), t).$$

Theorem 1 requires that the entries in X_0 are Bernstein, i.e. the variables $\{X_{0,j}\}$ are assumed to have sub-exponential tail behaviour. Consider now normalized design. We normalize each column X_j in X by dividing it by $\hat{\sigma}_j$, where $\hat{\sigma}_j^2 := \hat{\Sigma}_{j,j}$. Let thus $\tilde{X}_j := X_j/\hat{\sigma}_j$, $j = 1, \ldots, p$, $\tilde{X} := (\tilde{X}_1, \ldots, \tilde{X}_p)$ and

$$\hat{R} := \tilde{X}^T \tilde{X}/n.$$

Consider the normalized empirical compatibility constant

$$\tilde{\phi}^2(L, S) := \min\{ s u^T \hat{R} u :\ \|u_S\|_1 = 1,\ \|u_{-S}\|_1 \leq L \}.$$

Using the transfer principle as presented in [3] one can show lower bounds for $\tilde{\phi}^2(L,S)$ assuming only m-th order isotropy and no further moment conditions on the entries of X_0.

References

[1] G. Lecué and S. Mendelson *Compressed sensing under weak moment assumptions*, Arxiv preprint arXiv: 1401.2188 (2014).

[2] V. Koltchinskii and S. Mendelson, *Bounding the smallest singular value of a random matrix without concentration*, Arxiv preprint arXiv: 1312.3580 (2013). year=2013

[3] R.I. Oliveira *The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties*, Arxiv preprint arXiv: 13122903 (2013).

[4] S.A. van de Geer and A. Muro *Higher order isotropy and lower bounds for sparse quadratic forms* Arxiv preprint arXiv:1405.5995 (2014).

Reflections on quantum data hiding

Andreas Winter

We consider binary symmetric hypothesis testing of quantum states, ρ vs. σ, with equal prior probabilities. Even when focusing only on minimum error probability, this is a very rich problem when the measurements (POVMs) admissible are restricted in some way, for instance by conservation laws, bounded energy, etc [1, 2, 3]. Two very interesting restrictions, discussed in the talk, are:

(1) Locality in a composite system (LOCC);
(2) Linear operations and classical processing in continuous variable systems (GOCC).

Each such restriction $\mathcal{M}$ gives rise to a distinguishability norm on states, generalizing the trace (1-) norm [3]. In particular, *(quantum) data hiding* refers to $\|\cdot\|_{\mathcal{M}}$ being possibly much smaller than $\|\cdot\|_1$. Quantum data hiding, originally invented to show a limitation on LOCC in distinguishing globally orthogonal states, is actually a phenomenon arising generically in statistics whenever comparing a 'strong' set of measurements (i.e. decision rules) with a 'weak' one. The classical statistical analogue of this would be secret sharing, in which two perfectly distinguishable multi-partite hypotheses appear to be indistinguishable when accessing only certain marginals. The quantum versions are richer in that for example LOCC and GOCC allow for state tomography, so the states cannot be come perfectly indistinguishable but only nearly so, and hence the question is one of efficiency. Indeed, there are examples of almost perfectly distinguishable states which under the constraint are almost indistinguishable:

$$\|\rho-\sigma\|_1 \geq 2-\epsilon, \quad \|\rho-\sigma\|_{\mathrm{LOCC}} \leq \epsilon,$$
$$\|\rho'-\sigma'\|_1 \geq 2-\epsilon, \quad \|\rho'-\sigma'\|_{\mathrm{GOCC}} \leq \epsilon.$$

Curiously, in the first case, ρ and σ can be created by LOCC, i.e. they are separable; in the second case they can be created by GOCC, in fact they are probabilistic mixtures of coherent states.

There are two efficiency questions here: (a) How small can ϵ be in relation to the dimension of the system? Results on this are found in [4, 3] for LOCC. (b) How much information can be hidden reliably? In [5, 6] it is shown that in a bipartite $d \times d$-system, asymptotically $\log d$ bits can be hidden and that this is asymptotically optimal.

While it is known that data hiding by separable states is possible (i.e. the state preparation can be done by LOCC), it is open whether the optimal information efficiency of (asymptotically) $\log d$ bits can be achieved by separable states, or the best scaling of the hiding quality. Another open question is about generalizing this to larger number of parties (cf. cryptographic secret sharing).

References

[1] A. S. Holevo. *J. Multivar. Analysis* **3**:337- (1973).
[2] C. W. Helstrom. *Quantum Detection and Estimation.* Academic Press, 1976.
[3] W. Matthews, S. Wehner, A. Winter. *Commun. Math. Phys.* **291**:813-843 (2009).
[4] B. M. Terhal, D. P. DiVincenzo, D. W. Leung. *Phys. Rev. Lett.* **86**:8507-8510 (2001).
[5] P. Hayden, D. Leung, P. W. Shor, A. Winter. *Commun. Math. Phys.* **250**:371-391 (2004).
[6] A. Winter. In preparation (2014).

Large density matrix estimation for quantum systems based on Pauli measurements

Harrison Zhou
(joint work with T. Tony Cai, Yazhen Wang, Ming Yuan)

Modern scientific studies often need to learn and engineer quantum systems. Examples include quantum computation, quantum information and quantum simulation [Nielsen and Chuang (2000) and Wang (2011, 2012)]. These studies in particular frontier research in quantum computation generate tremendous interest in and great demand on quantum tomography. A quantum system is described by its state, and the state is often characterized by a complex matrix on some Hilbert space. The matrix is called density matrix. According to quantum physics the dimension of the Hilbert space and the size of the density matrix usually grows exponentially with the size of the quantum system. For the study of a quantum system, it is important but very difficult to know its state. In practice we may infer the quantum state by performing measurements on the quantum system. Statistically it is to estimate the density matrix based on measurements performed on a large number of quantum systems which are identically prepared in the same quantum state. The quantum literature refers quantum state tomography to as the reconstruction of the quantum state based on measurements obtained from measuring identically prepared quantum systems. Traditionally quantum tomography employs classical statistical models and methods to deduce quantum states from quantum measurements. Due to complexity of the problem, often times these approaches are neither very efficient nor effective from the statistical or computational point of view. A recent breakthrough establishes a

deep relationship between quantum tomography and compressed sensing (Gross et. al. (2010) and Wang (2013)). Compressed sensing develops innovative data acquisition techniques, efficient reconstruction methods and fast computational algorithms for recovering sparse signals and images from highly under-sampled observations [see Donoho (2006)]. It turns out that the density matrix reconstruction for a quantum state can be essentially recast as the matrix completion problem, which studies the reconstruction of low rank matrices based on under-sampled observations. Various methods are recently proposed for recovering low rank matrices by minimizing the squared residual sum plus some penalty, and the penalties used include nuclear-norm penalty [Candes and Plan (2009, 2011), Koltchinskii, Lounici and Tsybakov (2011) and Negahban and Wainwright (2011)], rank penalty [Bunea, She and Wegkamp (2011) and Klopp (2011)], the von Neumann entropy penalty [Koltchinskii (2011)], and the Schattenp quasi- norm penalty [Rohde and Tsybakov (2011)].

This paper considers the problem of density matrix estimation for a quantum spin system based on Pauli measurements. specifically we describe a quantum spin system by the d-dimensional complex space C^d and its quantum state by a complex matrix on C^d. From the theory of quantum physics, when measuring the quantum system by performing measurements on some observables which are Hermitian matrices, we obtain the measurement outcomes for each observable, where the measurements are random taking values from all eigenvalues of the observable, with the probability of observing a particular eigenvalue equal to the trace of the product of the density matrix and the projection matrix onto the eigenspace corresponding to the eigenvalue. To handle the spin up and down states of particles in the quantum spin system, we usually employ widely known Pauli matrices as observables to perform measurements and obtain the so-called Pauli measurements. Since all Pauli matrices have 1 and -1 eigenvalues, Pauli measurements takes discrete values 1 and -1, and the resulted measurement distributions can be characterized by binomial distributions. Our goal is to estimate the density matrix by the Pauli measurements. According to quantum physics, the dimension d increases exponentially in the number of particles in the quantum system, and the size of the density matrix may be comparable to or exceed the sample size, so we need to put the density matrix estimation problem in the framework of high-dimensional statistics where both dimension and sample size are allowed to go to infinity. Since Pauli matrices form a basis for all Hermitian matrices, we assume that the density matrix has a sparse representation under the basis and then employ thresholding methodology to recover the density matrix based on the Pauli measurements. We investigate the convergence rates of the proposed density matrix estimator under spectral norm and Frobenius norm. We establish the minimax lower bounds for the density matrix estimation problem and show that the constructed density matrix estimator can achieve the minimax lower bound and thus optimal.

The following are the major results.

Theorem 1. B_j, $j = 2, \ldots, p$, has spectral decomposition

$$B_j = 1 \cdot Q_{j+} + (-1) \cdot Q_{j-},$$

with

$$tr\left(Q_{j+}\right) = tr\left(Q_{j-}\right) = d/2,$$

and

$$tr\left(B_i Q_{j+}\right) = -tr\left(B_i Q_{j-}\right) = \begin{cases} d/2, & i = j \\ 0, & i \neq j \end{cases},$$

which implies $\left\{ \frac{B_1}{\sqrt{d}}, \frac{B_2}{\sqrt{d}}, \ldots, \frac{B_p}{\sqrt{d}} \right\}$ forms an orthonormal basis.

We assume that

$$\Theta = \left\{ \beta : \|\beta\|_0 \leq k_{n,p} + 1 \right\}.$$

and consider a thresholding Procedure. We show that

Theorem 2.

$$\inf_{\hat{\rho}} \sup_{\rho \in \Theta} E \left\| \hat{\rho} - \rho \right\|_F^2 \asymp \frac{k_{n,p} \log p}{n},$$

under the assumption $k_{n,p} \leq p^v$ for some positive $v < 1$ and the right hand side is bounded.

Theorem 3.

$$\inf_{\hat{\rho}} \sup_{\rho \in \Theta} E \left\| \hat{\rho} - \rho \right\|_{spectral}^2 \asymp \frac{k_{n,p}^2 \log p}{np},$$

under the assumption $k_{n,p} \leq p^v$ for some $0 < v < 1/4$.

References

[1] Artiles, L., Gill, R., and Guta, M., *An invitation to quantum tomography*, J. Roy. Statist. Soc. **67** (2005), 109–134.

[2] Donoho, D. L. and Johnstone, I. M., *Ideal spatial adaptation via wavelet shrinkage*, Biometrika **81** (1994), 425–455.

Reporter: Madalin Guta

Participants

Andreas Andresen
Weierstrass-Institute for Applied Analysis and Stochastics
Mohrenstr. 39
10117 Berlin
GERMANY

Prof. Dr. Koenraad Audenaert
Department of Mathematics
Royal Holloway University of London
Egham Hill
Egham, Surrey TW20 0EX
UNITED KINGDOM

Prof. Dr. Gilles Blanchard
Fachbereich Mathematik
Universität Potsdam
Am Neuen Palais 10
14469 Potsdam
GERMANY

Prof. Dr. Robin Blume-Kohout
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185
UNITED STATES

Prof. Dr. Daniel Burgarth
Institute of Mathematics & Physics
Aberystwyth University
Physical Sciences Bldg.
Aberystwyth, Wales SY23 3BZ
UNITED KINGDOM

Prof. Dr. Cristina Butucea
Lab. d'Analyse et de Mathématiques Appl.
UFR Mathématiques
Université Paris-Est Marne-la-Vallée
5, Bd. Descartes, Champs sur Marne
77454 Marne-la-Vallée Cedex 2
FRANCE

Yuri Campbell
Max-Planck-Institut für Mathematik in den Naturwissenschaften
Inselstr. 22 - 26
04103 Leipzig
GERMANY

Prof. Dr. Ismael Castillo
Laboratoire de Probabilites et Modeles Aleatoires
Université Paris VII
175 rue du Chevaleret
75013 Paris Cedex
FRANCE

Prof. Dr. Jens Eisert
Institut für Theoretische Physik
Freie Universität Berlin
Arnimallee 14
14195 Berlin
GERMANY

Prof. Dr. Chris Ferrie
Center for Quantum Information & Control,
Physics and Astronomy
University of New Mexico
1919 Lomas Blvd. NE
Albuquerque NM 87131
UNITED STATES

Prof. Dr. Akio Fujiwara
Department of Mathematics
Graduate School of Science
Osaka University
Machikaneyama 1-1, Toyonaka
Osaka 560-0043
JAPAN

Prof. Dr. Richard D. Gill
Mathematisch Instituut
Universiteit Leiden
Postbus 9512
2300 RA Leiden
NETHERLANDS

Prof. Dr. Georgii K. Golubev
Centre de Mathématiques et d'Informatique
Université de Provence
39, Rue Joliot-Curie
13453 Marseille Cedex 13
FRANCE

Prof. Dr. Ion Grama
Centre Yves Coppens
Campus de Tohannic
P.O. Box 573
56017 Vannes
FRANCE

Prof. Dr. David Groß
Physikalisches Institut
Universität Freiburg
79104 Freiburg i. Br.
GERMANY

Prof. Dr. Madalin Guta
School of Mathematical Sciences
The University of Nottingham
University Park
Nottingham NG7 2RD
UNITED KINGDOM

Prof. Dr. Marc Hoffmann
CEREMADE
Université Paris Dauphine
Place du Marechal de Lattre de Tassigny
75775 Paris Cedex 16
FRANCE

Sebastian Holtz
Fachbereich Mathematik
Humboldt Universität Berlin
10099 Berlin
GERMANY

Jana Jankova
Seminar für Statistik
ETH Zürich; HG G 17
Rämistr. 101
8092 Zürich
SWITZERLAND

Prof. Dr. Arnold Janssen
Mathematisches Institut
Heinrich-Heine-Universität Düsseldorf
Universitätsstr. 1
40225 Düsseldorf
GERMANY

Prof. Dr. Anna Jencova
Mathematical Institute
Slovak Academy of Sciences
Stefanikova 49
814 73 Bratislava 1
SLOVAKIA

Prof. Dr. Alexander S. Kholevo
Steklov Mathematical Institute
Department of Probability Theory
ul. Gubkina 8
Moscow 119 991
RUSSIAN FEDERATION

Jukka Kiukas
Department of Mathematics
The University of Nottingham
Nottingham NG7 2RD
UNITED KINGDOM

Dr. Bartek T. Knapik
Department of Mathematics
VU University Amsterdam
De Boelelaan 1081
1081 HV Amsterdam
NETHERLANDS

Richard Küng
Physikalisches Institut
Universität Freiburg
79104 Freiburg i. Br.
GERMANY

Prof. Dr. Anne Leucht
Institut für Mathematische Stochastik
TU Braunschweig
Pockelsstraße 14
38106 Braunschweig
GERMANY

Matthew Levitt
Department of Mathematics
The University of Nottingham
Nottingham NG7 2RD
UNITED KINGDOM

Housen Li
Max-Planck-Institut
für Biophysikalische Chemie
Am Fassberg 11
37077 Göttingen
GERMANY

Katarzyna Macieszczak
Department of Mathematics
The University of Nottingham
Nottingham NG7 2RD
UNITED KINGDOM

Prof. Dr. Enno Mammen
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg
GERMANY

Prof. Dr. Keiji Matsumoto
Principles of Informatics Research
Division
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo 101-8430
JAPAN

Prof. Dr. Alexander Meister
Fachbereich Mathematik
Universität Rostock
18051 Rostock
GERMANY

Dr. Thomas Monz
Institut für Experimentalphysik
Universität Innsbruck
Technikerstr. 25/4
6020 Innsbruck
AUSTRIA

Zacharie Naulet
8 B Rue Mademoiselle
75015 Paris
FRANCE

Prof. Dr. Natalie Neumeyer
Department Mathematik
Universität Hamburg
20146 Hamburg
GERMANY

Prof. Dr. Richard Nickl
Statistical Laboratory
Centre for Mathematical Sciences
Wilberforce Road
Cambridge CB3 OWB
UNITED KINGDOM

Prof. Dr. Michael Nussbaum
Department of Mathematics
Cornell University
Malott Hall
Ithaca, NY 14853-4201
UNITED STATES

Adélaide Olivier
École Nationale de la Statistique
et de l'Administration Economique
ENSAE CREST - Lab. de Statistique
3, Ave. Pierre Larousse
92245 Malakoff
FRANCE

Prof. Dr. Markus Reiß
Institut für Mathematik
Humboldt-Universität Berlin
Unter den Linden 6
10117 Berlin
GERMANY

Prof. Dr. Angelika Rohde
Ruhr-Universität Bochum
Fakultät für Mathematik
Lehrstuhl für Stochastik
44780 Bochum
GERMANY

Dr. Johannes Schmidt-Hieber
Mathematical Institute
University of Leiden
Niels Bohrweg 1
2300 RA Leiden
NETHERLANDS

Dr. Jiangwei Shang
Centre for Quantum Technologies
National University of Singapore
3 Science Drive 2
Singapore 117 543
SINGAPORE

Prof. Dr. Vladimir G. Spokoiny
Weierstrass-Institute for Applied
Analysis and Stochastics
Mohrenstr. 39
10117 Berlin
GERMANY

Adrian Steffens
Institut für Theoretische Physik
Freie Universität Berlin
Arnimallee 14
14195 Berlin
GERMANY

Dr. Arleta Szkola
Max-Planck-Institut für Mathematik
in den Naturwissenschaften
04103 Leipzig
GERMANY

Dr. Kristan Temme
IQIM
California Institute of Technology
1200 E. California Blvd.
91125 Pasadena
UNITED STATES

Prof. Dr. Sara van de Geer
Seminar für Statistik
ETH Zürich; HG G 17
Rämistr. 101
8092 Zürich
SWITZERLAND

Martin Wahl
Lehrstuhl für Statistik
Abteilung Volkswirtschaftslehre
Universität Mannheim
L7, 3-5
68131 Mannheim
GERMANY

Prof. Dr. Andreas Winter
Departament de Física
Àrea de Física Teòrica
Edifici C
Campus de la UAB
08193 Bellaterra (Barcelona)
SPAIN

Mayya Zhilova
Weierstraß-Institut für
Angewandte Analysis und Stochastik
Mohrenstr. 39
10117 Berlin
GERMANY

Prof. Dr. Huibin Zhou
Department of Statistics
Yale University
P.O.Box 208290
New Haven, CT 06520-8290
UNITED STATES