Journées Statistiques du Sud 2008

Conference Program

June 16-18 2008 – INSA Toulouse

The **Journées Statistiques du Sud** are a new series of workshops taking place in the sunny cities of Barcelona, Marseille, Montpellier, Nice and Toulouse. Centered on mini-courses and talks given by outstanding researchers, these workshops aim at providing a stimulating forum for exchanging ideas and learning about the latest developments on selected topics of intensive research in statistics. The 2007 edition took place in Nice. The **2008 edition** of Journées Statistiques du Sud will take place at the **INSA of Toulouse** located near the Université Paul Sabatier Campus. The topics chosen for the 2008 edition are **Statistics in High Dimensions** and **Statistics and Information Theory**.

Mini-courses

- László Györfi
  Empirical log-optimal portofolio selections
- Sara Van de Geer
  M-estimation and complexity regularization
- Alexander Tsybakov
  Sparse estimation in high-dimensional models

Invited speakers

- Sylvain Arlot
  V-fold cross-validation improved: V-fold penalization
- Arnak Dalalyan
  Estimation of the Effective Dimension Reduction Subspace
- Eustasio del Barrio
  Trimming methods in model checking
- Aurélien Garivier
  Identifying a Context Tree: BIC Estimator and Algorithm Context
- Christophe Giraud
  Inferring gene regulation networks
- Jean-Michel Marin
  Selection of Gaussian graphical models

The abstracts of the mini-courses and invited talks are collected at the end of this document. Up to date information is always available on the conference web site[1]

`http://djalil.chafai.net/jsds/`

[1] formerly `www(dot)math(dot)univ(dash)toulouse(dot)fr/(tilde)chafai/`

List of participants

- Akakpo, Nathalie, Paris XI
- Arlot, Sylvain, Paris XI
- Autin, Florent, Aix-Marseille I (LATP)
- Azaïs, Jean-Marc, Toulouse III (IMT)
- Baccini, Alain, Toulouse III (IMT)
- Baraud, Yannick, Nice
- Berthet, Philippe, Rennes I (IRMAR)
- Besse, Philippe, Toulouse III (INSA et IMT)
- Biau, Gérard, Paris VI (LSTA)
- Boitard, Simon, INRA (Génétique Cellulaire)
- Bontemps, Dominique, Paris XI
- Cavalier, Laurent, Aix-Marseille I (LATP)
- Celisse, Alain, Paris (AgroParisTech)
- Chafai, Djalil, Toulouse (INRA et IMT)
- Cornec, Matthieu, Paris X
- Dalalyan, Arnak, Paris VI (LPMA)
- Del Barrio, Eustasio, Valladolid
- Delmas, Céline, Toulouse (INRA SAGA)
- Dong, Qian, Rennes (ÉNS Cachan et IRMAR)
- Dupuy, Jean François, Toulouse III (IMT)
- Fort, Jean-Claude, Toulouse III (IMT)
- Gaiffas, Stéphane, Paris VI (LSTA)
- Gannaz, Irène, Grenoble (INP LJK)
- Garel, Bernard, Toulouse (ENSEEIHT et IMT)
- Garivier, Aurélien, Paris (CNRS et ENST)
- Girard, Robin, Grenoble (LJK)
- Giraud, Christophe, Jouy-en-Josas (INRA MIA)
- Györfi, László, Budapest
- Hebiri, Mohamed, Paris VII (LPMA)
- Huet, Sylvie, Jouy-en-Josas (INRA MIA)
- Klein, Thierry, Toulouse III (IMT)
- Lagnoux Renaudie, Toulouse III (IMT)
- Laurent, Béatrice, Toulouse (INSA et IMT)
- Lecué, Guillaume, Marseille (CNRS et LATP)
- Lerasle, Matthieu, Toulouse (INSA et IMT)
- Loubes, Jean-Michel, Toulouse III (IMT)
- Marin, Jean-Michel, INRIA Saclay et Paris XI
- Marteau, Clément , Aix-Marseille I (LATP)
- Mercier, Sabine, Toulouse (IMT, LSB)
- Meziani, Katia, Paris VII (LPMA)
- Michel, Bertrand, Paris XI

- Milhem, Hélène, Toulouse (INSA et IMT)
- Pelletier, Bruno, Montpellier II
- Pouet, Christophe, Aix-Marseille I (LATP)
- Prieur, Clémentine, Toulouse (INSA et IMT)
- Rabier, Charles-Elie, Toulouse III
- Robelin, David, Toulouse (INRA LGC/SAGA)
- Roquain, Etienne, Vrije Universiteit
- Ruiz-Gazen, Anne , Toulouse I (Gremaq et IMT)
- Salmon, Joseph, Paris VII (LPMA)
- Tsybakov, Alexandre, Paris (ENSAE ParisTech)
- Van de Geer, Sara, Zürich
- Verzelen, Nicolas, Paris XI
- Villa-Vialaneix , Nathalie, Toulouse (IMT)
- Villers, Fanny, Jouy-en-Josas (INRA)
- Willer, Thomas, Aix-Marseille I (LATP)
- Yao, Anne-Françoise, Aix-Marseille II (LMGEM)

Accomodations

Please book directly your room. You may find hotels suggestions on the conference web site. It is convenient to book a room in a hotel close to the Metro line B. Beware that in June, the hotels in Toulouse are often fully booked. Also, we recommend that you book your room as soon as possible (see conference website).

Access to the INSA

The INSA is near the Université Paul Sabatier Campus. This campus is located in the south west of the city. You can take a look at the map of the Campus (more precisely, page 2, rows 1 and 2 and column B). The campus is accessible by the Metro line B via two stations: Université Paul Sabatier and Faculté de Pharmacie (nearest to the INSA).

- From the city center, the INSA is accessible by the Metro Line B (nearest station is Faculté de Pharmacie)
- From the Toulouse-Blagnac airport (north east of the city), the INSA is accessible...
  - by taxi (about 30mn, 25-35 Euros)
  - or by the Shuttle which connects the airport to the Metro line B (at Compans-Caffarelli station)
- From the Matabiau train station, the INSA is accessible via Metro line A followed by Metro line B (line switch at Jean Jaures station).

Scientific committee

Y. Baraud (Nice), G. Biau (Paris 6), L. Cavalier (Aix-Marseille), B. Laurent-Bonneau (Toulouse), J.-M. Loubes (Toulouse), G. Lugosi (Barcelona).

Organizing committee

D. Chafaï, B. Laurent-Bonneau, M. Lerasle, J.-M. Loubes.

Administrative contact

Insatransfert - SAIC
Mme Patricia Jarry
135, avenue de Rangueil
31077 Toulouse Cedex 4 - France
mailto:patricia.jarry[at]insa-toulouse.fr

Sponsors

Institut National des Sciences Appliques (INSA) de Toulouse, Société de Mathématiques Appliquées et Industrielles (SMAI), Centre National de la Recherche Scientifique (CNRS), Région Midi-Pyrénées, Institut de Mathématiques de Toulouse (IMT), Université Paul Sabatier de Toulouse (UPS).

Journées Statistiques du Sud 2008
INSA Toulouse, June, 16-18

SCHEDULE

MONDAY, 16.06

9h00-9h45	Inscription
9h45-10h	Welcome
10h00-11h30	**Empirical log-optimal portofolio selection I (László Györfi)**
11h30-12h30	**Identifying a Context Tree : BIC Estimator and Algorithm Context (Aurélien Garivier)**

Lunch

14h00-15h30	**Sparse estimation in high-dimensional models I (Alexandre Tsybakov)**
15h30-16h00	Break
16h00-17h30	**M-estimation and complexity regularization I (Sara Van de Geer)**

TUESDAY, 17.06

9h00-10h30	**Sparse estimation in high-dimensional models II (Alexandre Tsybakov)**
10h30-11h00	Break
11h00-12h30	**M-estimation and complexity regularization II (Sara Van de Geer)**

Lunch

14h00-15h00	**Estimation of the Effective Dimension Reduction Subspace (Arnak Dalalyan)**
15h00-16h00	**Inferring gene regulation networks (Christophe Giraud)**
16h00-16h30	Break
16h30-17h30	**Selection of Gaussian graphical models (Jean-Michel Marin)**

Social Dinner

WEDNESDAY, 18.06

9h00-10h30	**Empirical log-optimal portofolio selection II (László Györfi)**
10h30-11h00	Break
11h00-12h00	**V-fold cross-validation improved : V-fold penalization (Sylvain Arlot)**
12h00-13h00	**Trimming methods in model checking (Eustasio Del Barrio)**

Lunch

Empirical log-optimal portfolio selections Abstract

László Györfi
Department of Computer Science and Information Theory
Budapest University of Technology and Economics,
Magyar Tudósok körútja 2.,Budapest, Hungary, H-1117
gyorfi@szit.bme.hu

Consider a market consisting of d assets. The evolution of the market in time is represented by a sequence of price vectors $\mathbf{s}_1, \mathbf{s}_2, \ldots \in \mathbb{R}_+^d$, where

$$\mathbf{s}_n = (s_n^{(1)}, \ldots, s_n^{(d)})$$

such that the j-th component $s_n^{(j)}$ of $\mathbf{s}_n$ denotes the price of the j-th asset on the n-th trading period. In order to normalize, put $s_0^{(j)} = 1$. $\{\mathbf{s}_n\}$ has exponential trend:

$$s_n^{(j)} = e^{nW_n^{(j)}} \approx e^{nW^{(j)}},$$

with average growth rate (average yield)

$$W_n^{(j)} := \frac{1}{n} \ln s_n^{(j)}$$

and with asymptotic average growth rate

$$W^{(j)} := \lim_{n\to\infty} \frac{1}{n} \ln s_n^{(j)}.$$

In order to apply the usual prediction techniques for time series analysis one has to transform the sequence price vectors $\{\mathbf{s}_n\}$ into a more or less stationary sequence of return vectors $\{\mathbf{x}_n\}$ as follows:

$$\mathbf{x}_n = (x_n^{(1)}, \ldots, x_n^{(d)})$$

such that

$$x_n^{(j)} = \frac{s_n^{(j)}}{s_{n-1}^{(j)}}.$$

Thus, the j-th component $x_n^{(j)}$ of the return vector $\mathbf{x}_n$ denotes the amount obtained after investing a unit capital in the j-th asset on the n-th trading period.

The *static portfolio selection* is a single period investment strategy. A portfolio vector is denoted by $\mathbf{b} = (b^{(1)}, \dots b^{(d)})$. The j-th component $b^{(j)}$ of $\mathbf{b}$ denotes the proportion of the investor's capital invested in asset j. We assume that the portfolio vector $\mathbf{b}$ has nonnegative components sum up to 1, that means that short selling is not permitted. The set of portfolio vectors is denoted by

$$\Delta_d = \left\{ \mathbf{b} = (b^{(1)}, \dots, b^{(d)});\, b^{(j)} \geq 0,\, \sum_{j=1}^{d} b^{(j)} = 1 \right\}.$$

For static portfolio selection, at time $n = 0$ we distribute the initial capital S_0 according to a fix portfolio vector $\mathbf{b}$, i.e., if S_n denotes the wealth at the trading period n, then

$$S_n = S_0 \sum_{j=1}^{d} b^{(j)} s_n^{(j)}.$$

One can show that

$$W := \lim_{n\to\infty} \frac{1}{n} \ln S_n = \lim_{n\to\infty} \max_j \frac{1}{n} \ln s_n^{(j)} = \max_j W^{(j)}.$$

Thus, any static portfolio selection achieves the maximal growth rate $\max_j W^{(j)}$.

One can achieve even higher growth rate for long run investments, if the tuning of the portfolio is allowed dynamically trading period after trading period. The *dynamic portfolio selection* is a multi-period investment strategy, where at the beginning of each trading period we rearrange the wealth among the assets. A representative example of the dynamic portfolio selection is the *constantly rebalanced portfolio (CRP)*, where we fix a portfolio vector $\mathbf{b} \in \Delta_d$, i.e., we are concerned with a hypothetical investor who neither consumes nor deposits new cash into his portfolio, but reinvest his portfolio each trading period. Note that in this case the investor has to rebalance his portfolio after each trading day to "corrigate" the daily price shifts of the invested stocks.

Let S_0 denote the investor's initial capital. Then at the beginning of the first trading period $S_0 b^{(j)}$ is invested into asset j, and it results in return

$S_0 b^{(j)} x_1^{(j)}$, therefore at the end of the first trading period the investor's wealth becomes

$$S_1 = S_0 \sum_{j=1}^{d} b^{(j)} x_1^{(j)} = S_0 \langle \mathbf{b}\,,\, \mathbf{x}_1 \rangle\,,$$

where $\langle \cdot\,,\, \cdot \rangle$ denotes inner product. For the second trading period, S_1 is the new initial capital

$$S_2 = S_1 \cdot \langle \mathbf{b}\,,\, \mathbf{x}_2 \rangle = S_0 \cdot \langle \mathbf{b}\,,\, \mathbf{x}_1 \rangle \cdot \langle \mathbf{b}\,,\, \mathbf{x}_2 \rangle\,.$$

By induction, for the trading period n the initial capital is S_{n-1}, therefore

$$S_n = S_{n-1} \langle \mathbf{b}\,,\, \mathbf{x}_n \rangle = S_0 \prod_{i=1}^{n} \langle \mathbf{b}\,,\, \mathbf{x}_i \rangle\,.$$

The asymptotic average growth rate of this portfolio selection is

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \ln S_n &= \lim_{n\to\infty} \left(\frac{1}{n} \ln S_0 + \frac{1}{n} \sum_{i=1}^{n} \ln \langle \mathbf{b}\,,\, \mathbf{x}_i \rangle \right) \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \ln \langle \mathbf{b}\,,\, \mathbf{x}_i \rangle\,, \end{aligned}$$

therefore without loss of generality one can assume in the sequel that the initial capital $S_0 = 1$.

If the market process $\{\mathbf{X}_i\}$ is memoryless, i.e., it is a sequence of independent and identically distributed (i.i.d.) random return vectors then we show that the best constantly rebalanced portfolio (BCRP) is the log-optimal portfolio:

$$\mathbf{b}^* := \arg\max_{\mathbf{b}\in\Delta_d} \mathbb{E}\{\ln \langle \mathbf{b}\,,\, \mathbf{X}_1 \rangle\}.$$

This optimality means that if $S_n^* = S_n(\mathbf{b}^*)$ denotes the capital after day n achieved by a log-optimum portfolio strategy $\mathbf{b}^*$, then for any portfolio strategy $\mathbf{b}$ with finite $\mathbb{E}\{(\ln \langle \mathbf{b}\,,\, \mathbf{X}_1 \rangle)^2\}$ and with capital $S_n = S_n(\mathbf{b})$ and for any memoryless market process $\{\mathbf{X}_n\}_{-\infty}^{\infty}$,

$$\lim_{n\to\infty} \frac{1}{n} \ln S_n \le \lim_{n\to\infty} \frac{1}{n} \ln S_n^* \quad \text{almost surely}$$

and maximal asymptotic average growth rate is

$$\lim_{n\to\infty} \frac{1}{n} \ln S_n^* = W^* := \mathbb{E}\{\ln \langle \mathbf{b}^*\,,\, \mathbf{X}_1 \rangle\} \quad \text{almost surely.}$$

We show several *examples* for constantly rebalanced portfolio.

In order to decrease the computational complexity of log-optimal portfolio we introduce the *semi-log-optimal portfolio*, where the function $\ln z$ is replaced by its second order Taylor expansion.

For a *general dynamic portfolio selection*, the portfolio vector may depend on the past data. Let $\mathbf{b} = \mathbf{b}_1$ be the portfolio vector for the first trading period. For initial capital S_0, we get that

$$S_1 = S_0 \cdot \langle \mathbf{b}_1 \, , \, \mathbf{x}_1 \rangle \, .$$

For the second trading period, S_1 is new initial capital, the portfolio vector is $\mathbf{b}_2 = \mathbf{b}(\mathbf{x}_1)$, and

$$S_2 = S_0 \cdot \langle \mathbf{b}_1 \, , \, \mathbf{x}_1 \rangle \cdot \langle \mathbf{b}(\mathbf{x}_1) \, , \, \mathbf{x}_2 \rangle \, .$$

For the nth trading period, a portfolio vector is $\mathbf{b}_n = \mathbf{b}(\mathbf{x}_1, \ldots, \mathbf{x}_{n-1}) = \mathbf{b}(\mathbf{x}_1^{n-1})$ and

$$S_n = S_0 \prod_{i=1}^{n} \left\langle \mathbf{b}(\mathbf{x}_1^{i-1}) \, , \, \mathbf{x}_i \right\rangle = S_0 e^{n W_n(\mathbf{B})}$$

with the average growth rate

$$W_n(\mathbf{B}) = \frac{1}{n} \sum_{i=1}^{n} \ln \left\langle \mathbf{b}(\mathbf{x}_1^{i-1}) \, , \, \mathbf{x}_i \right\rangle .$$

The fundamental limits reveal that the so-called *log-optimum portfolio* $\mathbf{B}^* = \{\mathbf{b}^*(\cdot)\}$ is the best possible choice. More precisely, on trading period n let $\mathbf{b}^*(\cdot)$ be such that

$$\mathbb{E} \left\{ \ln \left\langle \mathbf{b}^*(\mathbf{X}_1^{n-1}) \, , \, \mathbf{X}_n \right\rangle \middle| \, \mathbf{X}_1^{n-1} \right\} = \max_{\mathbf{b}(\cdot)} \mathbb{E} \left\{ \ln \left\langle \mathbf{b}(\mathbf{X}_1^{n-1}) \, , \, \mathbf{X}_n \right\rangle \middle| \, \mathbf{X}_1^{n-1} \right\} .$$

If $S_n^* = S_n(\mathbf{B}^*)$ denotes the capital achieved by a log-optimum portfolio strategy $\mathbf{B}^*$, after n trading periods, then for any other investment strategy $\mathbf{B}$ with capital $S_n = S_n(\mathbf{B})$ and with

$$\sup_n \mathbb{E} \left\{ (\ln \left\langle \mathbf{b}_n(\mathbf{X}_1^{n-1}) \, , \, \mathbf{X}_n \right\rangle)^2 \right\} < \infty,$$

and for any stationary and ergodic process $\{\mathbf{X}_n\}_{-\infty}^{\infty}$,

$$\limsup_{n \to \infty} \frac{1}{n} \ln \frac{S_n}{S_n^*} \leq 0 \quad \text{almost surely}$$

and

$$\lim_{n\to\infty}\frac{1}{n}\ln S_n^* = W^* \quad \text{almost surely,}$$

where

$$W^* := \mathbb{E}\left\{\max_{\mathbf{b}(\cdot)}\mathbb{E}\left\{\ln\left\langle \mathbf{b}(\mathbf{X}_{-\infty}^{-1})\,,\,\mathbf{X}_0\right\rangle \middle| \mathbf{X}_{-\infty}^{-1}\right\}\right\}$$

is the maximal possible growth rate of any investment strategy.

An empirical (data driven) portfolio strategy $\mathbf{B}$ is called *universally consistent* with respect to a class $\mathcal{C}$ of stationary and ergodic processes $\{\mathbf{X}_n\}_{-\infty}^{\infty}$, if for each process in the class,

$$\lim_{n\to\infty}\frac{1}{n}\ln S_n(\mathbf{B}) = W^* \quad \text{almost surely.}$$

For a fixed integer $k > 0$ large enough, let's apply the following approximation:

$$\mathbb{E}\{\ln\left\langle \mathbf{b}(\mathbf{X}_1^{n-1})\,,\,\mathbf{X}_n\right\rangle \mid \mathbf{X}_1^{n-1}\} \approx \mathbb{E}\{\ln\left\langle \mathbf{b}(\mathbf{X}_{n-k}^{n-1})\,,\,\mathbf{X}_n\right\rangle \mid \mathbf{X}_{n-k}^{n-1}\}$$

and

$$\mathbf{b}^*(\mathbf{X}_1^{n-1}) \approx \mathbf{b}_k(\mathbf{X}_{n-k}^{n-1}) = \arg\max_{\mathbf{b}(\cdot)}\mathbb{E}\{\ln\left\langle \mathbf{b}(\mathbf{X}_{n-k}^{n-1})\,,\,\mathbf{X}_n\right\rangle \mid \mathbf{X}_{n-k}^{n-1}\}.$$

Because of stationarity

$$\mathbf{b}_k(\mathbf{x}_1^k) \quad = \quad \arg\max_{\mathbf{b}}\mathbb{E}\{\ln\langle \mathbf{b}\,,\,\mathbf{X}_{k+1}\rangle \mid \mathbf{X}_1^k = \mathbf{x}_1^k\},$$

which is the maximization of the regression function

$$m_{\mathbf{b}}(\mathbf{x}_1^k) = \mathbb{E}\{\ln\langle \mathbf{b}\,,\,\mathbf{X}_{k+1}\rangle \mid \mathbf{X}_1^k = \mathbf{x}_1^k\}.$$

Thus, a possible way for asymptotically optimal empirical portfolio selection is that, based on the past data, sequentially estimate the regression function $m_{\mathbf{b}}(\mathbf{x}_1^k)$, and choose the portfolio vector, which maximizes the regression function estimate.

Next briefly summarize the basics of *nonparametric regression function estimation.*

Introduce the *kernel-based portfolio selection* strategies. Define an infinite array of portfolio selections $\mathbf{B}^{(k,\ell)} = \{\mathbf{b}^{(k,\ell)}(\cdot)\}$, where k, ℓ are positive

integers. For fixed positive integers k, ℓ, choose the radius $r_{k,\ell} > 0$ such that for any fixed k,

$$\lim_{\ell \to \infty} r_{k,\ell} = 0.$$

Then, for $n > k+1$, define the expert $\mathbf{b}^{(k,\ell)}$ by

$$\mathbf{b}^{(k,\ell)}(\mathbf{x}_1^{n-1}) = \arg\max_{\mathbf{b} \in \Delta_d} \sum_{\left\{k<i<n : \|\mathbf{x}_{i-k}^{i-1} - \mathbf{x}_{n-k}^{n-1}\| \le r_{k,\ell}\right\}} \ln \langle \mathbf{b}\,,\, \mathbf{x}_i \rangle \ ,$$

if the sum is non-void, and $\mathbf{b}_0 = (1/d, \ldots, 1/d)$ otherwise.

The good, data dependent choice of k and ℓ is doable borrowing current techniques from *machine learning*. In machine learning setup k and ℓ are considered as parameters of the estimates, called experts. The basic idea of machine learning is the combination of the experts, where an expert has large weight if its past performance is good. Combine the elementary portfolio strategies $\mathbf{B}^{(k,\ell)} = \{\mathbf{b}_n^{(k,\ell)}\}$ as follows: let $\{q_{k,\ell}\}$ be a probability distribution on the set of all pairs (k,ℓ) such that for all k, ℓ, $q_{k,\ell} > 0$. The combined strategy $\mathbf{B}$ arises from weighting the elementary portfolio strategies $\mathbf{B}^{(k,\ell)} = \{\mathbf{b}_n^{(k,\ell)}\}$ such that the investor's capital becomes

$$S_n(\mathbf{B}) = \sum_{k,\ell} q_{k,\ell} S_n(\mathbf{B}^{(k,\ell)}).$$

We prove that the portfolio scheme $\mathbf{B}$ is *universally consistent* with respect to the class of all ergodic processes such that $\mathbb{E}\{|\ln X^{(j)}|\} < \infty$, for $j = 1, 2, \ldots d$.

We present some *numerical results* obtained by applying the kernel based log-optimal algorithm to a *NYSE data set* from
www.szit.bme.hu/~oti/portfolio .

Sparse estimation in high-dimensional models

Alexandre Tsybakov

CREST and University of Paris 6

The aim of this short course is to give an introduction to statistical estimation in high-dimensional models (where the dimension p of the vector of unknown parameters is larger than the sample size n) under sparsity scenario. The model is called sparse if the number of non-zero coordinates of the vector of unknown parameters is much smaller than p. The quality of sparse estimation is usually assessed in terms of *model selection consistency* (i.e., recovering of the set of non-zero coordinates) and *sparsity oracle inequalities* (SOI) for the prediction risk. One of the most important issues is to build methods that attain optimal performances with respect to these two criteria under minimal assumptions on the dictionary (for example, in linear regression, this requirement is translated as minimal assumptions on the design matrix X). Sparse statistical estimation is closely related to the problem of compressive sensing in approximation theory, but is more complex because the noise is added. It is also related to the problem of aggregation of estimators since, using sparse estimation methods obeying the SOI, we can construct aggregates that are simultaneously optimal for convex, linear and model selection type aggregation.

First, an overview of the most popular methods of sparse statistical estimation will be given. They are mainly of the two types. Some of them, like the BIC, enjoy nice theoretical properties without any assumption on the dictionary but are computationally infeasible starting from relatively modest dimensions p. Others, like the Lasso or the Dantzig selector, are easily realizable for very large p but their theoretical performance is conditioned by severe restrictions on the dictionary. We will discuss and compare various types of such restrictions emphasizing that the Lasso and the Dantzig selector can be studied by similar methods and exhibit similar behavior.

We will then focus on *Sparse Exponential Weighting*, a new method of sparse recovery in regression, density and classification models realizing a compromise between theoretical properties and computational efficiency. The theoretical performance of the method is comparable with that of the BIC in terms of SOI for the prediction risk. No assumption on the dictionary is required, except for the standard normalization. At the same time, the

method is computationally feasible for relatively large dimensions p. It is constructed using the exponential weighting with suitably chosen priors, and its analysis is based on the PAC-Bayesian ideas in statistical learning. We will develop a general technique to derive sparsity oracle inequalities from the PAC-Bayesian bounds.

Bibiliography

Bickel, P.J., Ritov, Y. and Tsybakov, A.B. (2008) Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, to appear:
`http://www.imstat.org/aos/future_papers.html`

Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007a) Aggregation for Gaussian regression. *Annals of Statistics*, v.35, 1674-1697.

Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007b) Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, v.1, 169-194.

Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007c) Sparse density estimation with ℓ_1 penalties. *COLT-2007*, 530-543.

Dalalyan, A. and Tsybakov, A.B. (2008) Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, published on-line:
`http://dx.doi.org/10.1007/s10994-008-5051-0`

Juditsky, A., Rigollet, P. and Tsybakov, A.B. (2008) Learning by mirror averaging. *Annals of Statistics*, to appear: `http://www.imstat.org/aos/future_papers.html`.

M-estimation and complexity regularization

Sara van de Geer
Seminar für Statistik, ETH Zürich

1 Empirical processes

Consider a sample $Z_1, \ldots, Z_n$ of independent random variables, in some space $\mathcal{Z}$, and let $\gamma : \mathcal{Z} \to \mathbf{R}$ be a (measurable) function. We write the empirical average as

$$P_n\gamma := \frac{1}{n}\sum_{i=1}^{n}\gamma(Z_i),$$

and the theoretical mean as

$$P\gamma := \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\gamma(Z_i).$$

Let Γ be a collection of functions on $\mathcal{Z}$. Empirical process theory is about the study of quantities of the type

$$\mathbf{Z} := \sup_{\gamma\in\Gamma}|(P_n - P)\gamma|,$$

in particular the study of probability and moment inequalities for $\mathbf{Z}$. Of further interest is the empirical process

$$\nu_n := \{\nu_n(\gamma) := \sqrt{n}(P_n - P)\gamma : \ \gamma \in \Gamma\}.$$

Here, asymptotic continuity (tightness) is a key concept. This is the following property:

$$\sup_{\sigma(\gamma-\gamma_0)\le\epsilon_n}|\nu_n(\gamma - \gamma_0)| \to^{\mathbf{P}} 0,$$

as $n \to \infty$, with $\{\epsilon_n\}$ a sequence of positive numbers decreasing to zero. Moreover,

$$\sigma^2(\gamma) := \frac{1}{n}\sum_{i=1}^{n}\mathrm{var}(\gamma(Z_i)).$$

In fact, we will examine the *increments* or *modulus of continuity* of the empirical process, which is the behavior of, for instance, the moments

$$\psi(\epsilon) := \mathbf{E}\sup_{\sigma(\gamma-\gamma_0)\le\epsilon}|\nu_n(\gamma - \gamma_0)|,$$

as function of ϵ.

2 Application to M-estimation

Suppose $\Gamma \subset \mathbf{\Gamma}_0$ is a given collection of loss functions. The M-estimator is

$$\hat{\gamma} := \arg\min_{\gamma \in \Gamma} P_n \gamma.$$

It is to be understood as an estimator of the target

$$\gamma_0 := \arg\min_{\gamma \in \mathbf{\Gamma}_0} P\gamma.$$

Note since $\mathbf{\Gamma}_0 \supset \Gamma$, the target γ_0 may not be an element of the class Γ over which we perform empirical risk minimization. The best approximation within Γ of the target is defined as

$$\gamma^* := \arg\min_{\gamma \in \Gamma} P\gamma.$$

The excess risk is defined as

$$\mathcal{E}(\gamma) := P(\gamma - \gamma_0), \ \gamma \in \Gamma.$$

We moreover call $\mathcal{E}^* := \mathcal{E}(\gamma^*)$ the approximation error. The behavior of the excess risk $\hat{\mathcal{E}} := \mathcal{E}(\hat{\gamma})$ of the estimator $\hat{\gamma}$ will be our topic of interest. The following simple inequality is our starting point.

Lemma 2.1 *It holds that*

$$\hat{\mathcal{E}} \leq -\nu_n(\hat{\gamma} - \gamma^*)/\sqrt{n} + \mathcal{E}^*.$$

Thus, the excess risk $\hat{\mathcal{E}}$ is bounded by two terms. The second term is the approximation error, and the first term can be thought of as the estimation error. This first term can be handled using empirical process theory.

For example, suppose we can show that

$$\mathbf{V}_n := \sup_{\gamma \in \Gamma} \frac{|\nu_n(\gamma - \gamma^*)}{\psi(\sigma_\gamma \vee \sigma^*)}$$

is a tight sequence of random variables (for example that moments exist and do not explode as $n \to \infty$). Here, we define

$$\sigma_\gamma := \sigma(\gamma - \gamma_0),$$

and $\sigma^* := \sigma_{\gamma^*}$. Moreover, ψ is some (concave) strictly increasing function.

Lemma 2.2 *Suppose that the margin condition*

$$\mathcal{E}(\gamma) \geq G(\sigma_\gamma), \forall \ \gamma \in \Gamma$$

holds. Here, G is some (convex) increasing function. Assume that $G_\psi := G \circ \psi^{-1}$ is strictly convex. Let H be the convex conjugate of G_ψ. Then for all $0 < \delta < 1$,

$$(1-\delta)\hat{\mathcal{E}} \leq \delta H\left(\frac{\mathbf{V}_n}{\delta\sqrt{n}}\right) + (1+\delta)\mathcal{E}^*.$$

Thus, Lemma 2.2 gives a bound for the estimation error in terms of the modulus of continuity ψ of the empirical process.

3 Modulus of continuity and entropy

Definition *Let (Λ, d) be a subset of a metric space. The* δ-covering number *$N(\delta, \Lambda, d)$ of Λ is the smallest value of N such that there exist $\{\lambda_j\}_{j=1}^N$ with*

$$\min_{1 \leq j \leq N} d(\lambda, \lambda_j) \leq \delta, \ \forall \ \lambda \in \Lambda.$$

The entropy *$\mathcal{H}(\cdot, \Lambda, d)$ is then defined as*

$$\mathcal{H}(\cdots, \Lambda, d) := \log(1 + N(\cdot, \Lambda, d)).$$

For a probability measure Q on $\mathcal{Z}$, let $\|\cdot\|_Q$ denote the $L_2(Q)$-norm. Suppose the entropy condition

$$\sup_{\text{probability measures } Q} \mathcal{H}(\cdot, \Lambda, \|\cdot\|_Q) \leq \mathcal{H}(\cdot),$$

where $\mathcal{H}$ is a continuous function for which the integral

$$\psi(\cdot) := 24 \int_0^{\cdot} \sqrt{\mathcal{H}(u)} du,$$

exists.

We will prove the following theorem.

Theorem 3.1 *Assume that the functions γ in Γ are bounded in sup-norm by some constant K:*

$$\sup_{z \in \mathcal{Z}} |\gamma(z)| \leq K, \ \forall \ \gamma \in \Gamma.$$

Let H be the convex conjugate of $v \mapsto (\psi^{-1}(v))^2$. Then for $\epsilon^2 \geq 2H(4K/\sqrt{n})$, we have

$$\mathbf{E}\left(\sup_{\sigma(\gamma - \gamma_0) \leq \epsilon} |\nu_n(\gamma - \gamma_0)| \right) \leq \psi(4\epsilon)$$

4 Further themes

The technical tools we shall use involve Hoeffding's and Bernstein's inequalities, and contraction inequalities (Ledoux and Talagrand [1991]).

Note that Theorem 3.1 is a statement about the mean of the empirical process. In the part on empirical process theory, we will also discuss concentration inequalities, which say that the empirical process is concentrated around its mean with large probability (Bousquet [2002], Massart [2000]).

Lemma 2.2, shows that a good choice of the model class Γ involves a trade-off between estimation error and approximation error. We will discuss penalized empirical risk minimization and the so-called oracle inequalities (del Barrio

et al. [2007]). In particular, we will look at high-dimensional (generalized) linear models (van de Geer [2006], van de Geer [2007]),and ℓ_q penalties ($0 \leq q \leq 1$), and at additive models involving many components.

For most of the results, we will provide a complete proof. And of course, we will discuss many examples.

References

O. Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes rendus-Mathématique*, 334(6):495–500, 2002.

E. del Barrio, P. Deheuvels, and S. van de Geer. *Lectures on empirical processes: theory and statistical applications*. European Mathematical Society, Zürich, 2007.

M. Ledoux and M. Talagrand. *Probability in Banach spaces*. Springer-Verlag New York, 1991.

P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000.

S.A. van de Geer. High dimensional generalized linear models and the Lasso. *Research report, to appear in Annals of Statistics*, 2006.

S.A. van de Geer. The deterministic Lasso. *JSM proceedings*, 2007.

V-fold cross-validation improved: V-fold penalization.

Sylvain Arlot

Laboratoire de Mathmatiques
Universit Paris-Sud - Btiment 425
F-91405 ORSAY, FRANCE
(e-mail: sylvain.arlot@math.u-psud.fr)

We investigate the efficiency of V-fold cross-validation (VFCV) for model selection from the non-asymptotic viewpoint, and suggest an improvement on it, which we call "V-fold penalization".

First, considering a particular (though simple) regression problem, we will show that VFCV with a bounded V is suboptimal for model selection. The main reason for this is that VFCV "overpenalizes" all the more that V is large. Hence, asymptotic optimality requires V to go to infinity. However, when the signal-to-noise ratio is low, it appears that overpenalizing is necessary, so that the optimal V is not always the larger one, despite of the variability issue. This is confirmed by some simulated data.

In order to improve on the prediction performance of VFCV, we propose a new model selection procedure, called "V-fold penalization" (penVF). It is a V-fold subsampling version of Efron's bootstrap penalties, so that it has the same computational cost as VFCV, while being more flexible. In a heteroscedastic regression framework, assuming the models to have a particular structure, penVF is proven to satisfy a non-asymptotic oracle inequality with a leading constant almost one. In particular, this implies adaptivity to the smoothness of the regression function, even with a highly heteroscedastic noise. Moreover, it is easy to overpenalize with penVF, independently from the V parameter. As shown by a simulation study, this results in a significant improvement on VFCV in several non-asymptotic situations.

Estimation of the Effective Dimension Reduction Subspace

Arnak Dalalyan

Laboratoire de Probabilités et Modèles Aléatoires
Université Paris 6
4, Place Jussieu B. P. 188 75252 Paris Cédex 05, FRANCE
(e-mail: dalalyan@ccr.jussieu.fr)

The aim of this talk is to introduce a new procedure providing an estimator of the effective dimension reduction (EDR) subspace in the multi-index regression model with deterministic design and additive noise. More specifically, the problem of estimating the projection matrix $\Pi^* = \Theta\,\Theta^\top$ based on the observations $(x_1, Y_1), \dots, (x_n, Y_n)$ coming from the model

$$Y_i = f(x_i) + \varepsilon_i = g(\Theta^\top x_i) + \varepsilon_i, \qquad i = 1, \dots, n,$$

is addressed. In the general setup we are interested in, the covariates $x_i \in \mathbb{R}^d$, Θ is a $d \times m^*$ orthogonal matrix ($\Theta^\top\Theta = I_{m^*}$) and $g : \mathbb{R}^{m^*} \to \mathbb{R}$ is an unknown function. To be able to estimate Π^* consistently, we assume that $\mathcal{S}^* = Im(\Theta)$ is the smallest subspace satisfying $f(x_i) = f(\Pi_{\mathcal{S}} x_i)$, $\forall i = 1, \dots, n$, where $\Pi_{\mathcal{S}}$ stands for the orthogonal projector in $\mathbb{R}^d$ onto the subspace $\mathcal{S}$. We will focus our attention on the case where m^* is known.

Many methods dealing with the estimation of the EDR subspace perform principal component analysis on a family of vectors, say $\hat\beta_1, \dots, \hat\beta_L$, nearly lying in the EDR subspace. This is in particular the case for the structure-adaptive approach proposed by Hristache, Juditsky, Polzehl and Spokoiny (*Ann. Statist.* 2001). In contrast with this approach, we propose to estimate the projector onto the EDR subspace by the solution to the optimization problem

$$\text{minimize} \quad \max_{\ell=1,\dots,L} \hat\beta_\ell^\top (I - A)\hat\beta_\ell \quad \text{subject to} \quad A \in \mathcal{A}_{m^*},$$

where $\mathcal{A}_{m^*}$ is the set of all symmetric matrices with eigenvalues in $[0, 1]$ and trace less than or equal to m^*, with m^* being the true structural dimension. Under mild assumptions, $\sqrt{n}$-consistency of the proposed procedure is proved (up to a logarithmic factor) in the case when the structural dimension is not larger than 4. Moreover, the stochastic error of the estimator of the projector onto the EDR subspace is shown to depend on L logarithmically. This enables us to use a large number of vectors $\hat\beta_\ell$ for estimating the EDR subspace. The empirical behavior of the algorithm is studied through numerical simulations.

Trimming methods in model checking

Eustasio del Barrio

Departamento de Estadística, Universidad de Valladolid,
Prado de la Magdalena S/N,
47005 Valladolid, SPAIN
(e-mail: tasio@eio.uva.es)

This talk introduces an analysis of similarity of distributions based on measuring some distance between trimmed distributions. Our main innovation is the use of the impartial trimming methodology, already considered in robust statistics, which we adapt to the setup of model checking. By considering trimmed probability measures we introduce a way to test whether the *core* of the random generator underlying the data fits a given pattern. Instead of simply removing mass at non-central zones for providing some robustness to the similarity analysis, we develop a data-driven trimming method aimed at maximizing similarity between distributions. Dissimilarity is then measured in terms of the distance between the optimally trimmed distributions. Our main choice for appplications is the Wasserstein metric, but other distances might be of interest for different applications. We provide illustrative examples showing the improvements over previous approaches and give the relevant asymptotic results to justify the use of this methodology in applications.
Keywords. Trimmed distributions, similarity, transportation cost, asymptotics.

References

[1] Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A. and Matrán, C. (2008). Trimmed comparison of distributions. *To appear in J. Amer. Stat. Assoc.*

Identifying a Context Tree: BIC Estimator and Algorithm Context.

Aurélien Garivier

CNRS
ENST Paris
(e-mail: garivier@telecom-paristech.fr)

Stochastic chains with memory of variable length constitute a class of processes including Markov Chains, but potentially much more parsimonious. The idea behind the notion of variable memory models is that the probabilistic definition of each symbol only depends on a finite part of the past and the length of this relevant portion is a function of the past called "context". The set of all contexts satisfies the suffix property which means that no context is a proper suffix of another context. This property allows to represent the set of all contexts as a rooted labeled tree. With this representation the process is described by the tree of all contexts and a associated family of probability measures on the alphabet, indexed by the tree of contexts. Given a context, its associated probability measure gives the probability of the next symbol for any past having this context as a suffix. The pair composed by the context tree and the associated family of probability measures is called a probabilistic context tree. Originally also called finite memory source by Rissanen, this class of models recently became popular in the statistics literature under the name of Variable Length Markov Chains (VLMC) after an article by Buhlmann and Wyner. In 1983, Rissanen not only introduced the notion of variable memory models but he also proposed the algorithm Context to estimate the probabilistic context tree. The way the algorithm Context works can be summarized as follows. Given a sample produced by a chain with variable memory, we start with a maximal tree of candidate contexts for the sample. The branches of this first tree are then pruned until we obtain a minimal tree of contexts well adapted to the sample. We associate to each context an estimated probability transition defined as the proportion of time the context appears in the sample followed by each one of the symbols in the alphabet. Several variants of the algorithm Context have been presented in the literature: in all the variants the decision to prune a branch is taken by considering a gain function. A branch is pruned if the gain function assumes a value smaller than a given threshold. The estimated context tree is the smallest tree satisfying this condition. The estimated family of probability transitions is the one associated to the minimal tree of contexts. In his seminal paper Rissanen proved the weak consistency of the algorithm Context in the

case where the contexts have a bounded length, i. e. where the tree of contexts is finite. Buhlmann proved the weak consistency of the algorithm also in the finite case without assuming a prior known bound on the maximal length of the memory but using a bound allowed to grow with the size of the sample. In both papers the gain function is defined using the log likelihood ratio test to compare to candidate trees. On the other hand, Csiszar and Talata introduced a different approach for the estimation of the probabilistic context tree using the Bayesian Information Criterion (BIC). The BIC context tree estimator belongs to the family of penalized likelihood methods, which appear to be computationally efficient thanks to an elegant greedy procedure, the context tree maximizing algorithm. They proved strong consistency, but provided no finite-time control on the probability of over- or under-estimation. In this talk I will present non-asymptotic upper-bounds on the probability of error for the BIC estimator and for the Context algorithm. These bounds improve preceeding results by Galves and Maume, and require no hypotheses on the probability measures associated with each context. Their proof is made possible by the derivation of refined deviation inequalities for self-normalized martingales, which I shall briefly present.

Inferring gene regulation networks

Christophe Giraud[1,2]

1. Université de Nice - Sophia Antipolis,
Parc Valrose,
06108 Nice, France

2. INRA, Laboratoire MIA,
Domaine de Vilvert,
78352 Jouy-en-Josas, France

e-mail: christophe.giraud@unice.fr

Abstract

A current challenge in system biology is to infer the regulation network of a family of p genes from a n-sample of microarrays, with n (much) smaller than p. Gaussian graphical models are simple models to describe these regulation networks. We propose a procedure that performs Gaussian graph estimation by model selection. We introduce a collection of candidate graphs and then select one of them by minimizing a penalized empirical risk. We pay a special attention to the maximum degree of the graphs that we can handle and asses the performance of the procedure in a non-asymptotic setting. The good theoretical properties of the procedure are confirmed on numerical examples.

Keywords. Gene Regulation Networks, Gaussian Graphs, Model Selection, Sparsity.

Biological systems involve complex networks of interactions between entities such as genes or proteins. These networks can be conveniently represented by a graph. Each vertices of the graph corresponds to a protein or a gene, and an edge between two vertices represents a direct interaction. For example, Figure 1 records 1948 (known) interactions between 1458 proteins of the yeast. Recent biotechnological tools enable to produce a huge amount of proteomic or transcriptomic data. One of the challenge of the post-genomic is to infer the functional interactions between the genes or the proteins from these data. The task is challenging for the statistician due to the very high-dimensional nature of the data. For example, microarrays measure the expression level of a few thousand genes (typically 4000) whereas the sample size n is no more than a few tens. Since the number of possibles interactions between p genes is $p(p-1)/2$ (nearly ten millions if $p = 4000$), it seems hopeless to try to infer these interactions from $n \approx 20$ microarrays. This task is actually possible (up to some extend) thanks to the sparsity of the interaction network.

Valuable tools for analyzing the network of interactions are the Gaussian Graphical Models. The vector of the expression levels of the p genes is modeled by a Gaussian variable in $\mathbf{R}^p$. The Gaussian graph then represents the conditional dependences between the coordinates. More precisely, if $X = (X_1, \ldots, X_p)$ represent the expression levels of the p genes, the graph has an edge between the genes i and j if and only if X_i *is not independent of* X_j *conditionally on the other variables*. The goal of the statistician is to infer these edges from a n-sample of the variables X. The edges correspond to the non-zero entries of the partial correlation matrix, so when the sample size n is larger than p, a possible algorithm to infer the edges is to threshold the inverse of the empirical covariance matrix. This strategy is no more possible when n is (much) smaller than p and several new algorithms have been proposed. Unfortunately, the real performance of these algorithms are mostly unknown: the few theoretical results are only valid under restrictive conditions on the covariance matrix and they assume that the sample size n tends to infinity.

We propose a new statistical procedure to estimate the graph of conditional dependences of X. We first introduce a collection of candidate graphs and then select one of them by minimizing a penalized empirical risk. The performance of the procedure is assessed in a non-asymptotic setting without any hypotheses on the covariance matrix. These good theoretical properties of

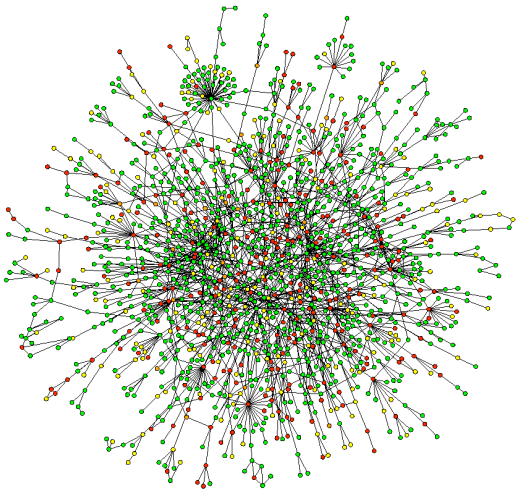

Figure 1: Protein-Protein interaction network of the yeast.

the procedure are confirmed by numerical results. Since we are interested on the maximal "size" of the graph that we can infer, we pay a special attention to the maximal degree D of the graphs that we can handle. This maximal degree turns to be roughly $n/(2\log(p/D))$, which means that p should stay small compared to $de^{n/(2d)-1}$, where d is the degree of the graph of conditional dependences.

Selection of Gaussian graphical models

Sophie Donnet[1], Jean-Michel Marin[2⋆]

[1] CEREMADE, Université Paris Dauphine, France
[2] INRIA Saclay, projet SELECT, Université Paris-Sud, France
⋆ jean-michel.marin@inria.fr

The last decade has witnessed the apparition of applied problems typified by very high-dimensional variables, in marketing database or gene expression studies for instance. Graphical models (Lauritzen (1996)) enable concise representations of associational relations between variables. If the graph is known, the parameters of the model are easily estimated. However, a quite challenging issue is the selection of the most appropriate graph for a given dataset. We consider this problem and the case of decomposable Gaussian graphical models (Dawid and Lauritzen (1993)).

Let $\mathcal{G} = (V, E)$ be an undirected graph with vertices $V = \{1, \ldots, p\}$ and set of edges E. We suppose that $\mathcal{G}$ is decomposable. To each vertex $v \in V$ of the graph, we associate a random variable y_v. Let $\mathbf{y} = (y_1, \ldots, y_p)$, a graphical model is a family of distributions on $\mathbf{y}$ which are Markov with respect to a graph. A Gaussian graphical model is such that

$$\mathbf{y}|\mathcal{G}, \Sigma_{\mathcal{G}} \sim \mathcal{N}_p\left(0_p, \Sigma_{\mathcal{G}}\right) , \tag{1}$$

where Σ_G is a positive definite matrix which ensures that the distribution of $\mathbf{y}$ is Markov with respect to $\mathcal{G}$. $\Sigma_{\mathcal{G}}$ ensures that the distribution of $\mathbf{y}$ is Markov iff $(i,j) \notin E \Longleftrightarrow \left(\Sigma_{\mathcal{G}}^{-1}\right)_{(i,j)} = 0$.
We observe a sample $\mathbf{y}^1, \ldots, \mathbf{y}^n$ from (1) (the data are centered). We would like to identify the set of most relevant graphs. For the considered multivariate random phenomenon, we are interested in the set of most relevant conditional independence structures.

We consider the Bayesian paradigm. Conditionally on $\mathcal{G}$, we use a Hyper-Inverse Wishart (HIW) distribution associated to the graph $\mathcal{G}$ as prior distribution on $\Sigma_{\mathcal{G}}$: $\Sigma_{\mathcal{G}}|\mathcal{G}, \delta_{\mathcal{G}}, \Phi_{\mathcal{G}} \sim \mathrm{HIW}_{\mathcal{G}}\left(\delta_{\mathcal{G}}, \Phi_{\mathcal{G}}\right)$ where $\delta_{\mathcal{G}} > 0$ and $\Phi_{\mathcal{G}}$ is a $p \times p$ symmetric positive definite matrix. Conditionally on $\mathcal{G}$, the HIW distribution is conjugate

$$\Sigma_{\mathcal{G}}|\mathcal{G}, \mathbf{y}^1, \ldots, \mathbf{y}^n, \delta_{\mathcal{G}}, \Phi_{\mathcal{G}} \sim \mathrm{HIW}_{\mathcal{G}}\left(\delta_{\mathcal{G}} + n, \Phi_{\mathcal{G}} + \sum_{i=1}^{n} \mathbf{y^i}\left(\mathbf{y^i}\right)^{\mathrm{T}}\right) . \tag{2}$$

Moreover, for such a prior, $f(\mathbf{y}|\mathcal{G}, \delta_{\mathcal{G}}, \Phi_{\mathcal{G}}) = \dfrac{h_{\mathcal{G}}(\delta_{\mathcal{G}}, \Phi_{\mathcal{G}})}{(2\pi)^{np/2} h_{\mathcal{G}}\left(\delta_{\mathcal{G}} + n, \Phi_{\mathcal{G}} + \sum_{i=1}^{n} \mathbf{y^i}\left(\mathbf{y^i}\right)^{\mathrm{T}}\right)}$ where $h_{\mathcal{G}}$ is the normalizing constant of the HIW distribution associated to the graph $\mathcal{G}$. Finally, we assume a uniform prior distribution in the space of graphs: $\pi(\mathcal{G}) \propto 1$. In that case,

$$\pi\left(\mathcal{G}|\mathbf{y}^1, \ldots, \mathbf{y}^n, \delta_{\mathcal{G}}, \Phi_{\mathcal{G}}\right) \propto f(\mathbf{y}|\mathcal{G}, \delta_{\mathcal{G}}, \Phi_{\mathcal{G}}) . \tag{3}$$

It is well-known that (3) is sensible to the specification of the hyper-parameters $\delta_{\mathcal{G}}$ and $\Phi_{\mathcal{G}}$ (Giudici and Green (1999), Jones et al. (2005)). In this work, we address this problem and present different strategies. Then, we introduce a new sampling scheme to explore the space of graphs and conclude with some experiments on simulated and real datasets.

Dawid, A. and Lauritzen, S. (1993) Hyper Markov Laws in the statistical analysis of decomposable graphical models, Annals of Statistics, 21, 1272-1317.

Giudici, P. and Green, P. (1999) Decomposable graphical Gaussian model determination, Biometrika, 86, 785-801.

Jones, B., Carvalho, C., Dobra, A., Hans, C. Carter, C. and West, M. (2005) Experiments in Stochastic Computation for High-Dimensional Graphical Models, Statistical Science, 20, 388-400.

Lauritzen, S. (1996) Graphical models, Oxford Statistical Science Series, 17, Oxford Science Publications