**Supplementary materials**

# Probability and stochastic processes theory

We include the basic probability theory definitions and properties of stochastic processes from the literature for easy reference (Capiński and Kopp 2004; Durrett 2019; Steele 2001; Hui-Hsiung 2006; Arnold 1974, 1995; Lehmann and Casella 2006).

**Def. 1.** Consider a randomized experiment as the construct upon an abstract probability space defined by the triad . A probability space includes the sample space  or set of all possible random outcomes, the sigma-algebra  or set of all possible events, and the probability measure  with . A *Random variable,*which we denote as , is the mathematical construct for experimental observables, a real-valued *measurable function*  with  (in ).

\*A real-valued *measurable function* is such that, for any Borel set  with  Borel sigma-algebra of , the inverse function  is contained in sigma-algebra  (measurable by probability ).

\*\*A Borel sigma-algebra  of  is the minimal sigma-algebra containing , the set of semi-bounded intervals  with .

**Def. 2.** *Cumulative distribution* of a *random variable* , denoted  as in the equation Eq. , is the probability  induced by inverse function  over  the semi-bounded intervals. *The probability density* of the *random variable*, denoted as , is the derivative, with respect to , of the *cumulative distribution* :



*\** Consideranother *random variable*. We may introduce the *joint* *cumulative distribution,* denoted by  as in the equation Eq. for a pair of *random variables* , as a real-valued *measurable vector function*  with  (in ). In other words, a *joint* *cumulative distribution* is a probability  induced by an inverse function  (or equivalently ) over  the semi-bounded rectangles. The *Joint probability density* of the pair of *random variables* , denoted as , is the mixed derivative with respect to  of the *cumulative distribution* .



\*\*The *Conditional cumulative distribution,* denoted as  in the equation , of a random variable  with respect to another random variable , is the probability  that is induced by  the subtraction of inverse function  from  over semi-bounded rectangles . *Conditional probability density* of random variable  with respect to  which we denote as  is a more technical definition that instead of the probability  considers  over a sufficiently narrow rectangle . Then  is just the derivative of  with respect to . In the probability theory literature  is also Radon-Nikodym derivative of the probability induced by inverse function  with respect to the probability induced by inverse function  over Borel sets .



\*\*\* Baye’s theorem for the *conditional cumulative distributions*  (or *probability densities* ) is a critical consequence of the equation Eq. that reads as the equation Eq. .



**Def. 3.** -th order moment  of a cumulative distribution  (or probability density ) is the expected value  Eq. of the random variable  to a natural number . The first-order moment  is the expected value (mean value) , and the second-order central moment  computed around  is the variance .



\*\*a vital assumption so that all the moments are finite for all . This assumption of finite moments need not cause concern, for all the random variables available in experimental practice appear to be strictly bounded  , so all the moments exist.

**Def. 4.** *Sample* is the collection of random outcomes  with index  counting up to a finite number  (sample size) of independent realizations of the randomized experiment . For a *random variable,* , this set is  , or simply  with the index  substituting .

\*We employ the sample  to determine statistics such as  the unbiased estimator -th order moment  Eq. . An estimator of the -th order moment  is the expected value  computed over a sample  with , and which distinguishes from  computed over  (or probability density ). The first-order moment estimator  is the sample mean .The second-order central moment , computed around , is the variance estimator .



**Def. 5.** *Random vector,* denoted as , (boldfaced letter) is a real-valued *measurable vector function*  that comprises entries  with  counting a finite collection of *random variables* and with .

\*A real-valued *measurable vector function* is such that, for any Borel set  with  Borel sigma-algebra of , the inverse function  is contained in sigma-algebra  (measurable by probability ).

\*\*A Borel sigma-algebra  of  is the minimal sigma-algebra containing  with , the set of -dimensional semi-bounded hyper-rectangles (-dimensional cartesian product of semi-bounded intervals) .

\*\*\*We may consider an essential generalization of a *random vector*, a *random field* , a collection of *random variables* (real-valued *measurable functions*)  with a vector index  over the 3-dimensional continuous spatial domain  , and with . The rigorous definition of  requires the extension of Borel sigma-algebra to the continuous function space that we do not include here. However, in neuroimaging practice, we are limited to a discrete spatial domain  with , and with index  counting  recording points. In these points, a neuroimage or random vector is defined as , a finite sub-collection from the *random field* . For simplicity, we use a short notation  when referring to the sub-collection.

**Def. 6.** *Cumulative distribution* of a *random vector* , denoted as  in the equation Eq. , is the probability  that is induced by inverse function  over semi-bounded hyper-rectangles . *The probability density* of a *random vector* denoted as  is the mixed derivative with respect to  of the *cumulative distribution* .



**Def. 7.**The definition of a *Vector stochastic process*, denoted as , employing a second-dimension index  that shall appear by default and which is different from the index  that counts the vector first-dimension, is the collection of *random vectors*  with  over the continuous time domain  and with .

\*An alternative definition of a *vector stochastic process* is the sub-collection  (in vector notation  with ) from a *dynamic random field* , collection of real-valued measurable functions  with  and .

\*\*The set of entries  of a *vector stochastic process* adjust to the nominal definition of *stochastic process*, i.e. for each entry  , the collection of real-valued *measurable functions*  with  over the continuous time domain  and with .

\*\*\*Note that defining a *vector stochastic process*  is very common in continuous time domain , but doing so with  suits much better to neuroimaging practice. Also note that a rigorous definition for  is through measurable functions over a Filtration, or a process of increasing sigma-algebras in time, so that the stochastic process represents a causal system.

**Def. 8.** *A Vector time series* which we denote as  is the countable (finite in neuroimaging practice) sub-collection from a *vector stochastic process*  over discrete time domain  with  and with index  counting  time points. In other words, *vector time series* which we denote  is a real-valued *measurable matrix function*  with .

\*For the sake of simplicity, we use the short notation  when referring to a *vector time series* . Hence, when referring to the physical time we calculate it as , increasing according to a sampling period . Note that we may also refer to  or  as a more general sub-collection with index  counting  time points that do not increase with a fixed sampling period. Also note that one could consider vector time series with an infinite number of time points as  but this would require *measurable functions* defined over an infinite codomain.

\*\*A real-valued *measurable matrix function* (analogously to previous *measurable vector function*) is such that, for any Borel set  with  Borel sigma-algebra of , the inverse function  is contained in sigma-algebra  (measurable by probability ).

\*\*\*Borel sigma-algebra  of  is (analogously to previous Borel sigma-algebra  of ) the minimal sigma-algebra containing  with , the set of all -dimensional semi-bounded hyper-rectangles (-dimensional cartesian product of -dimensional semi-bounded hyper-rectangles) .

**Def. 9.** *Cumulative distribution* of a *vector time series*  or simply sub-collection  (in other literature *finite-dimensional cumulative distribution* of a *vector stochastic process* ) which we denote as  Eq. is the probability   that is induced by inverse function  over semi-bounded hyper-rectangles . *Probability density* of a *vector time series*  (finite-dimensional *probability density* of a *vector stochastic process* ) which we denote as  is the mixed derivative with respect to  of the *cumulative distribution* .



\*We can now introduce the important condition of *stationarity*. A *vector stochastic process*  with  is *strictly stationary* if the *probability densities* (or equivalently *cumulative distributions*) of all possible sub-collections (not only sub-collections of *vector time series*)  with index  are translationally invariant, i.e. the identity Eq. holds for all .



\*\*Note that translational invariance Eq. also applies to all possible *vector time series*  with physical time  and index  which may be defined from the *vector stochastic process*  modifying the values of , and . A translationally invariant *probability density* (or equivalently *cumulative distribution*) then, is so that for each value , and  defining *vector time series*  the identity Eq. holds for all .



\*\*\*Stationarity is a strong condition but seems to be approximately valid for conditions valid during resting-state electrophysiological studies (Damoiseaux et al. 2006; Greicius 2008; Smith et al. 2013) or a task in a block design (Larson-Prior et al. 2013; Van Essen et al. 2013). It may be substituted by the assumption of *local stationarity*.

**Def. 10.** One may extend the concept of moments, from a *random variable*  to a *dynamic random field*, a collection of *random variables*  over the continuous spatial domain  and the time domain . This extension is the essential concept of the multivariate statistics designated *cumulant* , an -th moment computed over the type of sub-collection  with  counting a finite set of points in space  and time  from a *dynamic random field* .

\*The cumulant  may simply be understood as the -th moment that is computed over  with fixed , a sub-collection in time domain from a *vector stochastic process* , but yet this definition is too technical to be included here. However, the first-order cumulant  and second-order cumulant  are the mean value  and auto-covariance matrix  Eq. that are computed from the corresponding *cumulative distribution* (or *probability density*). In the notation of *vector time series,* the cumulant notation reads as . Thus, equation Eq. with  and with  reads as follows Eq. .





\*\*Assume the vector stochastic process  *strictly stationary*. Then, the *cumulative distributions*  and the cumulants  that are defined for a sub-collection , are translationally invariant. Thus, expressing the *cumulative distributions* or the *cumulants* we may arbitrarily omit one of the time points say  with . The cumulants are then just a function of  that we denote as subspace  Eq. with exclusion of  that leads to “”. In the notation of *vector time series* with  () and  we denote the subspace  and the equation Eq. reads as Eq. .





\*\*\* Another important property is that of mixing of a *vector stochastic process*  (or a *dynamic random field* ) indicating that the span of dependence between time points is small. A form of the mixing condition is integrability of the *cumulants* , for any , over the subspace  Eq. . The mixing condition also seems reasonable for electrophysiology since predictability of future states from the past is always limited. In the notation of *vector time series* integrable reads as summable over the subspace  Eq. .





**Def. 11.** *Trajectories* or sample path, or simply *realizations* of a *vector stochastic process* (also of any sub-collection in time or a *vector time series*) are the sample  (also  or ). This sample is, as usual, defined over the set of random outcomes  with  and with index  counting up to a number  (sample size) of independent realizations of the randomized experiment. The realizations are in short notation  (also  or ) with the index  substituting .

\*We may employ the samples of a sub-collection  or a *vector time series*  to determine statistics such as  (short notation for  with  fixed) or  the unbiased estimator of the -th order cumulant  or  that is also too technical to be included here. The sampled estimator of the first-order cumulant  is the mean value estimator , and the second-order cumulant  is the auto-covariance matrix estimator  Eq. . In the notation of *vector time series* these estimators Eq. read as Eq. .





# gamma-MAP and implementation of ssSBL

***Lemma1 (Andrews and Mallows)***

Let  be random variable distributing with the following probability density Eq. df:



where is a normalization constant. Then the following equality holds:

(4.2)

where is the Truncated Gamma pdf, with a lower truncation limit.

***Proof of Lemma1:***

The Normal/Laplace probability density function derived from the Gibbs model with Elastic Net penalization can be rearranged as:

(4.3)

Using the integral representation of the Gaussian scale mixtures based on the Andrews and Mallows lemma for the Laplace term, we can represent the formula [B3] above:

(4.4)

Alternatively, writing the Normal distribution explicitly in (A1-4), we obtain:

(4.5)

To further simplify the expression (B5), we can rearrange the term on the right by multiplying and dividing by :

(4.6)

Through using the following change of variables

we arrive at:

Then, using the definition for the Gamma probability density function truncated in the interval , denominated truncated Gamma density:

We can finally demonstrate that the Normal/Laplace probability density function can be represented as the following scaled mixture of Gaussians:

(4.7)

where .

***Vector form of Lemma1***

This can be extended for the case in which is the vector in the following expression, with a matrix function of the vector argument .

(4.8)

where is the L1 norm and is the L2 norm, for the complex-valued vector and the truncated gamma distribution upon the vector is expressed as follows:

■ (4.9)

## Extension of the Andrews and Mallows Lemma to the complex-valued hierarchical Elastic Net using measure-densities

The results for the hierarchical Elastic Net can be extended to the complex domain by modifying the Andrews and Mallows Lemma. For the complex-valued Elastic Net, the integral representation holds.

(4.10)

where the variances are defined as .

The measurable space in which the variable is defined as an unnormalized density function given by the Gaussian pdf and its variance is dependent on the random variable which has Truncated Gamma pdf.

(4.11)

Then, the measure in the space product of and is had density represented as an unnormalized product of Gaussian and Gamma densities.

(4.12)

## Structured space-frequency sparsity modes within the hierarchical complex-valued Elastic-Net

We introduce additional tensor group penalization for the Hierarchical Elastic Net on the 3D cartesian space of generators, samples, and frequencies with:

(4.13)

where , , ; refers to a specific Gray matter area ; refers to a specific frequency band . Then the transformed prior of the parameters is described analytically by the following distribution:

(4.14)

where and

In vector form they are expressed as:

(4.15)

where and , , . The unnormalized distribution upon is represented as:

where the -th element diagonal element are the variances. The full vector Bayesian model is as follows:

, , (4.16)

, , (4.17)

(4.18)

where and , ,

## Bayesian first type maximum a posteriori (parameters) analysis with the hierarchical complex-valued Elastic-Net prior

**Proposition1:**

For the joint distribution of data and parameters the following factorization holds, for simplicity, we avoid the use of argument for frequency and samples :

(4.19)

The quantities (posterior mean) and (posterior covariance) are defined as follows:

(4.20)

(4.21)

The quantity (ensemble covariance) is given by:

(4.22)

Note that we avoid using the frequency argument , used to define the model in (4.19], for the sake of the readability of the derivations.

**Proof of Proposition1**

This proposition can be demonstrated by writing their distributions explicitly.

The form of the resultant distribution can be found by analyzing the terms that depend on the parameters (exponential argument) in the formula above:

(4.23)

Reorganizing in (4.23] of terms 2 and 5 to render and in terms 3 and 4 to render

(4.24)

From (4.24] based on and we obtain

(4.25**)**

Completing (4.25] with the term

(4.26)

Completing (4.26] with the terms , and we obtain:

(4.27)

Then terms 3, 4, and 6 in (4.27] can be reorganized into since to obtain:

(4.28]

But combining terms 3 and 4 in (4.28] yields

(4.29)

From (4.29], it holds that:

(4.30)

## Bayesian second type maximum a posteriori (hyperparameters) analysis with the hierarchical complex-valued Elastic Net prior

From the joint distribution we can derive iteratively and approximated representation of the Type II-Likelihood :

(4.31)

(4.32)

where

(4.33)

The analysis of the previous section yields

(4.34)

Then iteratively upon fixed values of the posterior mean the Type II-Likelihood is expressed as:

(4.35)

Where the covariances and are

***Parameter estimators***

The parameters are determined in the previous iteration in terms of the iteratively linear source transfer operator

(4.36)

The source transfer operator is defined upon fixed values of the hyperparameters:

(4.37)

where

and

This source transfer operator simplifies the computations of expressed as:

(4.38)

where

The residuals are determined in the previous iteration in terms of the iteratively linear residual transfer operator

(4.39)

The residual transfer operator is defined upon fixed values of the hyperparameters and the source transfer operator :

(4.40)

This simplifies the computations of expressed as:

(4.41)

The estimation formulas can be derived by applying maximum a posterior of the combined Type II Likelihood and priors. To do so, we reformulate the targeted hyperparameters:

(4.50)

(4.51)

(4.52)

***Variances***

First the estimator of variances can be computed from the stationary values of the expression:

(4.53)

Due to the chain rule of matrix derivatives, the first term can be expressed in a close form.

(4.54)

where and

Yielding:

(4.55)

The second and third term derivatives are:

(4.56)

(4.57)

Substituting the derivatives, we obtain:

(4.58)

But since it holds that the expression is much more compact:

(4.59)

Using the auxiliary quantity so that we obtain:

(4.60)

Therefore, the only possible solution for the conditions set by the problem statement and estimator can be obtained from with:

(4.61)

(4.62)

***Regularization parameters and***

The estimator of the regularization parameters and can be computed from the stationary values of the expression below. The computation follows the same steps as for the variances. See section A5 **Proposition A5-IIc**:

(4.63)

Writing explicitly the distribution and using the chain rule in the derivative :

(4.64)

where ;

The derivative is given by:

(4.65]

where with shape and rate

From this, we obtain the equation for :

(4.66)

Due to the chain rule of matrix derivatives, the first term for the parameter can be expressed as:

(4.67)

The derivative is given by:

(4.68)

where is the number of active sources.

The derivative is given by:

(4.69)

where with shape and rate

From this, we obtain the equation for :

(4.70]

*or*

(4.71)

***Noise parameter***

The estimator of the noise parameter can be computed from the stationary values of the expression below, following the same steps as for the variances.

(4.72)

Due to the chain rule of matrix derivatives, the first term for the parameter can be expressed as:

(4.73)

where and with

The derivative is given by:

(4.74)

The derivative is given by:

(4.75)

where with shape and rate

From this, we obtain the equation for :

(4.76)

*or*

(4.77)

**Algorithm statistics:** The ssSBL allows screening out the neural space by thresholding the posterior distribution statistic: the ratio of the posterior mean and posterior variances. After convergence, the estimated source activity can be thresholded employing an unbiased statistic: this is due to the posterior distribution of source activity of formula [5.1], where the quantities (posterior mean) and (posterior covariance) are defined as follows:

(4.78)

(4.79)

In this distribution, is the posterior mean and the posterior covariance. The z-statistic for the analysis of variance has the following form:. A plausible way to screen out the active sources is to extract the set of nodes that return a value of the z-statistic greater than 1: .

**Implementation details:** The high computational cost for obtaining employing a matrix inversion operation can be avoided by using the economic singular value decomposition (SVD) of the lead field , and the Woodbury identity, leading to: .

The update formulas in Proposition A5-II a), b) are consistent with the sparsity constraint in both the ENET and ELASSO models, since the elements of the effective prior variance matrix (or equivalently **)** select which elements of become zero. When , the -th row and -th -column of the matrix tend to zero vectors, from where . In the same way, if some parameters are very small in a previous iteration (, i-th diagonal element), they will lead to in the next iteration (equations (A5-2) and (A5-8)). In some algorithms, this property usually means that if one activation is set to zero (e.g., removed from the active set) in an iteration, it will not appear as part of the solution. In our case, however, we do not prune to zero the small coefficients. Therefore, although unlikely, a “zeroed” activation might be re-estimated in a future iteration and contribute to the solution.

The nonlinear terms in (A5-5) and (A5-10) are obtained from the derivative of the normalization constants. These terms decrease strictly with respect to their arguments leading to smaller values of F for higher values of and , which is equivalent in both cases to have more zero elements in . The measurements variance in (A5-6) and (A5-7) is generally considered superfluous in the learning process, because it only acts as a scale factor for the parameters and usually decelerates the algorithm convergence. In our case, we fix it to , for all time points.

We also use fixed values for the parameters of the Gamma distribution of ENET’s and and ELASSO’s . In particular, we chose for ENET , which preserves the monotony of (C6-5) (in the sense that only one zero of exists) and , where is such that the prior is flexible: with a mean () and variance (). Following a similar flexible strategy with a mean () and variance (), we chose ENET’s , which is also in the same order of magnitude of the numerator in (C6-3), and , as a form of regularization in the denominator of (C6-3). This combination of similar and priors keeps adequate balance, allowing flexibility in our learning of different degrees of sparsity. Our choice of ELASSO’s is similar to that of the ENET’s , which preserves the monotony of (C6-10) (in the sense that only one zero of exists) and , where is such that the prior is flexible: with a mean () and variance ().

References

Arnold, Ludwig. 1974. “Stochastic Differential Equations: Theory and Application.”

———. 1995. “Random Dynamical Systems.” *Dynamical Systems*, 1–43.

Capiński, Marek, and Peter Ekkehard Kopp. 2004. *Measure, Integral and Probability*. Vol. 14. Springer.

Damoiseaux, Jessica S, SARB Rombouts, Frederik Barkhof, Philip Scheltens, Cornelis J Stam, Stephen M Smith, and Christian F Beckmann. 2006. “Consistent Resting-State Networks across Healthy Subjects.” *Proceedings of the National Academy of Sciences* 103 (37): 13848–53.

Durrett, Rick. 2019. *Probability: Theory and Examples*. Vol. 49. Cambridge university press.

Essen, David C. Van, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. 2013. “The WU-Minn Human Connectome Project: An Overview.” *NeuroImage* 80: 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041.

Greicius, Michael. 2008. “Resting-State Functional Connectivity in Neuropsychiatric Disorders.” *Current Opinion in Neurology* 21 (4): 424–30.

Hui-Hsiung, Kuo. 2006. *Introduction to Stochastic Integration*. Universitext. New York: Springer-Verlag. https://doi.org/10.1007/0-387-31057-6.

Larson-Prior, L. J., R. Oostenveld, S. Della Penna, G. Michalareas, F. Prior, A. Babajani-Feremi, J. M. Schoffelen, et al. 2013. “Adding Dynamics to the Human Connectome Project with MEG.” *NeuroImage* 80: 190–201. https://doi.org/10.1016/j.neuroimage.2013.05.056.

Lehmann, Erich L, and George Casella. 2006. *Theory of Point Estimation*. Springer Science & Business Media.

Smith, Stephen M, Diego Vidaurre, Christian F Beckmann, Matthew F Glasser, Mark Jenkinson, Karla L Miller, Thomas E Nichols, Emma C Robinson, Gholamreza Salimi-Khorshidi, and Mark W Woolrich. 2013. “Functional Connectomics from Resting-State FMRI.” *Trends in Cognitive Sciences* 17 (12): 666–82.

Steele, J Michael. 2001. *Stochastic Calculus and Financial Applications*. Vol. 1. Springer.