Annotated Bibliography

Suzie Brown

SEQUENTIAL MONTE CARLO

Gordon, Salmond, and Smith, 1993, "Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation"

Original reference for SMC.

Kitagawa, 1996, "Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models"

Nice introduction to SMC. Review of other nonlinear filtering techniques: extensions to Kalman filtering.

Del Moral, 2013, *Mean Field Simulation for Monte Carlo Integration*

Loads of rigorous results about SMC e.g. convergence, rates, CLTs.

Doucet and Johansen, 2011, "A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later"

Andrieu, Doucet, and Holenstein, 2010, "Particle Markov Chain Monte Carlo Methods"

Introduces particle MCMC methods, including particle Gibbs with conditional SMC.

RESAMPLING

Kitagawa, 1996, "Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models"

Comparison of multinomial, stratified & systematic resampling. And the effect of presorting. [in appendix]

Douc, Cappé, and Moulines, 2005, "Comparison of Resampling Schemes for Particle Filtering"

Comparison of Monte Carlo variance between mutli, res-multi, strat, syst. CLTs for resampled particles.

Lee, Murray, and Johansen, 2019, "Resampling in Conditional SMC Algorithms"

Implementation of low variance resampling within conditional SMC.

Fox, 2003, "Adapting the Sample Size in Particle Filters through KLD-Sampling"

Gandy and Lau, 2016, "The Chopthin Algorithm for Resampling"

Whitley, 1994, "A Genetic Algorithm Tutorial"

Carpenter, Clifford, and Fearnhead, 1999, "Improved Particle Filter for Nonlinear Problems"

Gerber, Chopin, and Whiteley, 2019, "Negative Association, Ordering and Convergence of Resampling Methods"

Del Moral, Doucet, and Jasra, 2012, "On Adaptive Resampling Strategies for Sequential Monte Carlo Methods"

MATRIX RESAMPLING

Webber, 2019, *Unifying Sequential Monte Carlo with Resampling Matrices*

- Allows number of particles to vary from one iteration to the next
- Allows resampling such that the weights after resampling are unequal
- Introduces the class of "matrix resampling schemes", and then the extension of this to encompass resampling schemes that cause the population size to vary randomly (but not unboundedly).
- To be representable by a matrix (i.e. in this class) a resampling scheme must satisfy the property that parental indices are conditionally independent given weights.
- Includes NO RESAMPLING as an example resampling scheme! The reason being that it actually is included in the class of matrix resampling schemes
- Also includes as an example the independent implementation of MVB resampling, where you don't enforce constant N so everything is stochastically rounded independently (in this paper this is called "Bernoulli resampling"). This is not in the class of matrix resampling schemes, but it is in the extended class (Defn 2.2).
- It is shown that all matrix resampling schemes are unbiased
- Under additional assumptions, it is also shown that such resampling schemes have properties: MC estimates converge; and the variance of these estimates is bounded above.
- The above results hold even if the resampling scheme to use at each iteration is chosen on-the-fly — notably including adaptive resampling, parallel resampling, adaptive pruning/enrichment, and you can imagine some other eccentric mixes of resampling schemes!
- *Complete* resampling schemes are defined as the subclass of matrix resampling schemes for which the weights after resampling are equal. (This is the class used by Li2020, although there they don't consider it a restriction from the more general class; this is just how they define matrix resampling in the first place.)
- Three comparisons are made between resampling schemes, in terms of conditional variance: multi $\geq$ strat; multi $\geq$ res-multi; multi $\geq$ res-strat. None of these examples are novel results, but perhaps the proof technique for strat could be useful as (at first glance anyway) it seems more understandable than that of DCM05.
- Theorem 3.1 proves that stratified resampling on particles sorted by a certain functional of the states is optimal within the matrix resampling class, in terms of the conditional variance used e.g. in DCM05. This generalises the result of Li2020 to higher dimensions, but proves optimality only in one sense. The functional is tailored to the test function φ appearing in the conditional variance, as one might expect, and it also cannot typically be computed.

- The authors suggest alternative sorting rules that give good (but of course not optimal) performance: Hilbert curve sorting as proposed in GCW19; or approximating the intractable functional by binning.
- Theorem 3.2: asymptotic error/CLT results for a variety of resampling schemes, with or without sorting by some coordinate.
- Recommendations: replace multi with strat wherever possible; and sort before resampling if there exists a suitably influential coordinate by which to sort.

Li et al., 2020, *Stratification and Optimal Resampling for Sequential Monte Carlo* (arXiv v2)

- Resampling is not assumed to preserve the number of particles
- Resampling schemes where the parental indices are conditionally independent (but not necessarily i.i.d.) — notably multi, strat, res-multi and res-strat — can be represented as a matrix conditional on the weights, where each element P_{ij} is the conditional probability of particle i choosing parent j (so the rows sum to 1, and column j sums to Nw_j).
- For multinomial resampling, all parental indices are conditionally i.i.d., so the rows of P_{ij} are identical, each containing the vector of weights.
- For stratified resampling, P is a "staircase matrix": see Figure2(b) in the paper for an illustration. Upon thinking a little it is clear why you end up with a staircase matrix.
- Residual resampling sets some rows to have a single 1 element and the rest 0 (these correspond to the deterministic assignments) and the remaining rows are constructed according to which scheme is used for the residuals, where the weights are now replaced with residual weights. See Fig2(c)(d) in the paper for an illustration
- Equation (1) gives an explicit expression for the conditional variance induced by resampling, the same metric used in DCM05. This expression, in the case of stratified resampling, seems much nicer than the expression used in DCM05, which could provide a more elegant proof that multi dominates strat. (EDIT: this expression no longer appears in arxiv v2, so possibly it wasn't correct)
- In Section 2.5 the authors admit that there is generally little to be gained by resampling such that the weights after resampling are unequal, even though I think it was one of the authors (Liu) who was proposing in previous works that this could be a useful strategy. (EDIT: this remark has been removed in v2)
- The authors prove that, in 1D, ordered stratified resampling is optimal (in the senses of conditional variance, expected squared energy distance, and earth-mover distance) among the class of resampling schemes considered. But their class only includes resampling schemes where the parental indices are conditionally independent. This excludes syst, res-syst, SSP,...
- Theorem 3 tightens the convergence rate bound for Hilbert-ordered stratified resampling from $O(N^{-1-1/d})$ (in GCW19) to $O(N^{-1-2/d})$.
- Proposition 2 verifies the conjecture of GCW19 that the Hilbert curve is an optimal way to order the particles in dimension $d > 1$

PARALLEL RESAMPLING

See also: Vergé, Dubarry, Del Moral and Moulines (2013) "On parallel implementation of sequential Monte Carlo methods: The island particle model".

Whiteley, Lee, and Heine, 2016, "On the Role of Interaction in Sequential Monte Carlo Algorithms"

- Introduces the αSMC framework, a generalisation encompassing e.g. SIS, bootstrap PF, adaptive resampling.
- Sections 3–4 present convergence results for generic αSMC.
- In Section 5.1 they use the αSMC framework to demonstrate that running multiple independent PFs and then averaging them is not a good solution to distributed PFing. The multiple independent PFs can be expressed as an instance of αSMC.
- "Degree of interaction" is represented by graph degree: SIS corresponds to the identity matrix or graph with self-loops only (degree= 1, the minimum possible); Bootstrap PF corresponds to the matrix with $1/N$ in every element, or the complete graph on N vertices with self-loops (degree= N, the maximum possible); adaptive resampling chooses either the SIS or bootstrap graph at each iteration.
- The structure of these graphs (see definition of "B-matrices" on p.514) is such that (1) each node has the same degree, (2) all self-loops are present, (3) all the connected components are complete subgraphs. These conditions are sufficient to ensure all the convergence properties of Theorem 2.
- Algorithm 4 is a procedure for choosing the interaction matrix on-the-fly. By construction the resulting matrix is a B-matrix. The algorithm partitions $[N]$ into interaction blocks, i.e. the connected components of the corresponding graph. The sooner the while loop terminates, the smaller the connected components and the less interaction.
- Section 5.4 contains simulation results. They demonstrate that the "random" and "greedy" variants of Algorithm 4 keep the level of interaction very low, the "simple" variant less so, but all much lower than the full interaction in the bootstrap PF.
- They also compare the MSE of the various techniques (Figure 4), seeming to show that adaptive resampling outperforms the others as long as its threshold is not too high (e.g. the usual $N/2$ threshold is good in this example).
- The authors' "random" and "greedy" methods do perform comparably to adaptive resampling though, and they seem to rarely require interaction between more than 2 or 4 particles at a time, so they might be more suitable for distributed settings?

Lee and Whiteley, 2016, "Forest Resampling for Distributed Sequential Monte Carlo"

- This paper extends the work of WLH16 on αSMC
- The idea is to find practical αSMC algorithms, where the on-the-fly choice of resampling type doesn't require too much particle interaction, so it really does work on-the-fly in parallel computing. In WLH16 they only considered removing interaction within the actual resampling step, not while choosing the resampler.
- The permissable matrix/graph representations of resampling are limited to B-matrices as in WLH16, corresponding to *disjoint unions of complete graphs*
- They find that the partial ordering on such graphs (merging some connected components to create a coarser graph) corresponds to an ordering on the respective ESS's (using an extended defn of ESS that holds on restrictions of $[N]$ and may be weighted by some numbers c)
- This suggests Algorithm 2, which recursively searches for a graph satisfying the convergence criterion of WLH16, coarsening the graph at each recursion
- They then drill down into the implementation detail, managing somehow to be simultaneously nitty-gritty and abstract.

- TBH I didn't understand a lot of this paper...

Murray, Lee, and Jacob, 2016, "Parallel Resampling in the Particle Filter"

- Motivates the need for parallelisable resampling schemes
- Introduces *Metropolis resampling*, a Metropolis approximation of Multinomial resampling, i.e. run a Markov chain targetting the Categorical distribution parametrised by the normalised weights.
- The Metropolis resampler is parallelisable because it doesn't require the weights to be normalised (i.e. their sum calculated) and it only compares weights pairwise (for the accept/reject step), so no operations require access to all the particles at once.
- The Metropolis resampler is biased, because the Markov chain is only run for a finite number of iterations, so it leads to biased likelihood estimates (so it's no use for pseudo-marginal-type PMCMC).
- The complexity of the Metropolis resampler is $O(NB)$ where B is the number of Metropolis steps used per iteration, which may scale as a function of N. B is a tuning parameter, trading off speed vs. accuracy; the authors provide some guidelines for choosing its value, which require that an upper bound on weights is available.
- Introduces *rejection resampling*, which uses rejection sampling to choose the parents. Only possible when an upper bound on the weights is known. Two variants: the initial proposal is uniform over $[N]$, or is deterministically equal to the offspring index.
- Runtime of rejection resampling for each particle is random — it keeps proposing until it gets an acceptance — so it's not as parallel as Metropolis resampling; some particles will take longer to resample and leave other processors idling.
- Rejection resampling is unbiased.
- If the deterministic first proposal is used, rejection sampling has lower variance than Multinomial resampling, although this gain decreases as the variance of weights increases. With the Uniform first proposal, it samples the parents exactly from the Categorical distribution.
- *Partial rejection resampling* can be used to reduce the expected runtime of rejection sampling, or in cases where no upper bound on weights is available but some almost-upper-bound V is known. Instead of resampling with the weights w_i, resample with $w_i \wedge V$, and remove the bias by setting the weights after resampling to the ratio of the true weight and the resampling weight. This means any particularly high-weight particles for which $w_i > V$ do not have their weights completely reset, just decreased by a factor of V (before normalisation: recall that parallel resampling avoids normalising weights). Taking V small reduces the runtime but also reduces the effectiveness of resampling by resetting fewer weights.
- Lots of simulation results showing how performance compares between multinomial, stratified, systematic, Metropolis and rejection resampling, and their dependence on the number of particles and variance of weights. In their example a tight analytic upper bound on weights is available; I don't know how realistic this is in general.

BACKWARD SIMULATION

Kitagawa, 1996, "Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models"

Some solutions to ancestral degeneracy: fixed lag smoother, forward-backward-type algorithm.

Doucet and Johansen, 2011, "A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later"

Lindsten and Schön, 2013, "Backward Simulation Methods for Monte Carlo Statistical Inference"

A whole book on backward simulation. [Chapter 5] describes backward simulation and ancestor sampling in particle MCMC.

Whiteley, 2010, "Discussion on 'Particle Markov Chain Monte Carlo Methods"'

In the discussion, Nick Whiteley introduces (remarkably briefly) the idea of ancestor sampling in particle Gibbs.

CONVERGENCE OF GENEALOGIES

Also consider looking at: Tavaré 1984 "Line-of-descent and genealogical processes, and their applications in population genetics models"; Donnelly 1991 "Weak convergence to a Markov chain with an entrance boundary: ancestral processes in population genetics"; Donnelly Tavaré 1995 "Coalescents and genealogical structure under neutrality"; Griffiths Tavaré 1994 "Sampling theory for neutral alleles in a varying environment"; Marjoram 1992 "Correlation structures in applied probability" (PhD thesis, UCL); Schweinsberg 2003 "Coalescent processes obtained from supercritical Galton-Watson processes"

Kingman, 1982, "On the Genealogy of Large Populations"

- Introduces the n-coalescent (in a very nice clear way) with the same notation we still use
- **Theorem:** suppose $\nu_{1:N}$ are exchangeable and independent across generations (i.e. neutral Cannings models, including neutral W-F for example) and $\mathrm{Var}[\nu_1] \to \sigma^2 \in (0, \infty)$ and $\mathbb{E}[\nu_1^m] \leq M_m$ for all $m \in \mathbb{N}$. Then the n-genealogies scaled by $\lfloor N\sigma^{-2}t \rfloor$ converge to the n-coalescent in the sense of FDDs.
- Condition $\sigma^2 > 0$ excludes e.g. the neutral Moran model.
- n-coalescent also applies for models where ν_j are not exchangeable or independent across generations, as long as the genealogies are Markov at least up to error $O(N^{-1})$ and the transitions satisfy $p_{\xi\eta} = q_{\xi\eta}\sigma^2 N^{-1} + o(N^{-1})$, where q's are transition probs of n-coalescent
- The n-coalescent is a good robust model for *large neutral* populations
- Genealogy decouples into a jump chain and a pure death process
- There exists the Kingman coalescent as infinite-dimensional embedding of the n-coalescents

Kingman, 1982, "The Coalescent"

Broadly, this paper introduces the Kingman coalescent (as opposed to n-coalescent) and proves some properties

Tavaré, 1984, "Line-of-Descent and Genealogical Processes, and Their Applications in Population Genetics Models"

- Review article on genealogical models and their application to population genetics

Möhle, 1998, "Robustness Results for the Coalescent"

Necessary & sufficient conditions for convergence of Cannings model to a coalescent process more general than Kingman. Allowing large mergers but not simultaneous mergers.

- Population size can vary over time, but only deterministically
- Individuals are not exchangeable, just assume the random assignment condition
- Offspring counts must be independent but not necessarily identically distributed across generations
- This means time scale must be allowed to vary over time: τ_N becomes $\tau_N(t)$ and c_N becomes $c_N(t)$, or $c(t)$ in Möhle's notation
- Only proves convergence of FDDs
- Conditions of theorem are still very strong, requiring infinitely many moments to be bounded. There is now a condition on one mixed moment that wasn't needed in Kingman1982.
- The conditions are sufficient but not necessary
- Time scale $\tau_N(t)$ is allowed to be chosen freely, but the theorem only holds when it is an appropriate function (i.e. an inverse of c_N similar to the usual) so this is not a very great generalisation over defining τ in the usual way.

Möhle and Sagitov, 1998, "A Characterization of Ancestral Limit Processes Arising in Haploid Population Genetics Models"

- Assume $\nu_{1:N}$ are exchangeable, and i.i.d. across generations, and the population size N is constant
- Necessary & sufficient conditions are given for (FDD) convergence of the genealogies to a Λ-coalescent, and the correct measure Λ is uniquely constructed from infinitely many moment limits. (Note: Möhle's notation uses μ for the measure rather than Λ.)
- The conditions are: (I) infinitely many pure moment limits exist, slightly different from the condition 2a of Möhle1998; (II) the exchangeable version of the mixed moment condition 2b of Möhle1998.
- Under the additional condition $c_N \to 0$ we also get weak convergence to the Λ-coalescent, although the proof is not explicit in this paper (it just refers to the methods of another work).

Sagitov, 1999, "The General Coalescent with Asynchronous Mergers of Ancestral Lines"

- Assume $\nu_{1:N}$ are exchangeable, and i.i.d. across generations, and the population size N is constant
- k-mergers (but not simultaneous mergers) are allowed
- 3 necessary conditions are given for FDD convergence to some limit of the form $p_{\xi\eta} = \delta_{\xi\eta} + V_N q_{\xi\eta} + o(V_N)$, where $Q = (q_{\xi\eta})$ is some Markov generator, and $V_N \to 0$.
- Additionally, a subset of the necessary conditions are shown to be sufficient for the above asymptotic relation to hold with specific Q corresponding to a Λ-coalescent, on the (constant, deterministic) time scale $T_N^{-1} \sim V_N$.
- Unfortunately I didn't really understand the second and third conditions...
- As one would expect, the necessary conditions can be shown to hold when Kingman's condition $\sup_N \mathbb{E}[\nu_1^k] < \infty \forall k \geq 2$ applies (see Remark 1)

Pitman, 1999, "Coalescents with Multiple Collisions"

- Defines Λ-coalescents in the notation still used, proves its existence.
- Notes how Kingman coalescent and Bolthausen-Sznitman coalescent (indeed the whole class of β-coalescents) can be seen as special cases
- Properties: exchangeable random partition; Exponential waiting times with parameter determined by Λ; criterion for coming down from infinity; restriction to $[n]$
- This paper studies the Λ-coalescents from the point of view of coalescent theory / exchangeable random partitions. It doesn't make any reference to population dynamics that might lead to such a limit. That angle was covered in Sagitov1999 (who independently discovered the Λ-coalescents).

Möhle, 1999, "Weak Convergence to the Coalescent in Neutral Population Models"

- Population size can vary over time, but only deterministically
- Offspring counts must be independent but not necessarily identically distributed across generations — this ensures the genealogical process is Markovian, but not time-homogeneous
- Individuals are not necessarily exchangeable, but our standing assumption (here called the "random assignment condition") is assumed
- Time scale $\tau_N(t)$ is allowed to be chosen freely, but the theorem only holds when it is an appropriate function (i.e. an inverse of c_N similar to the usual) so this is not a very great generalisation over defining τ in the usual way.
- Theorem conditions are exactly the same as in Möhle1998
- Convergence is now proved additionally in the weak sense, as opposed to just FDDs
- Proof technique is overall the one I used for weak convergence of SMC genealogies, except Möhle's doesn't have the external expectations. These expectations appear in my proof because the coupled process can only be defined conditionally on $\mathcal{F}_\infty$, because the generations are not independent

Möhle, 2000, "Total Variation Distances and Rates of Convergence for Ancestral Coalescent Processes in Exchangeable Population Models"

- Fixed population size N
- Offspring counts are exchangeable and i.i.d. across generations
- Treats full characterisation of possible limiting coalescents (I think the class is known as Ξ-coalescents), where there are possibly simultaneous and/or multiple mergers. The particular limiting coalescent can be characterised by (mixed) factorial moments, here denoted $\Phi_a(b_1, \ldots, b_a)$. These quantities are central to the analysis, and results are given in terms of them, so can be applied to any exchangeable model.
- It is shown that the Kingman coalescent appears as the limit if and only if $\lim \frac{\mathbb{E}[(\nu_1)_3]}{N\mathbb{E}[(\nu_1)_2]} = 0$. (See equation (14) or (16).) This is the exchangeable version of the main theorem condition I use in my SMC genealogies work. I think this work is the earliest use of this simplified condition.
- The above condition also implies that $c_N \to 0$ and $\frac{\mathbb{E}[(\nu_1)_2(\nu_2)_2]}{N^2\mathbb{E}[(\nu_1)_2]} = 0$.
- Total variation bounds are proved in the general case, left in terms of the particular transition probabilities of the model (see Thm 1).

- More specific bounds are calculated for the case where the model is in the domain of attraction of Kingman's coalescent (i.e. second moments dominate third moments, as above / in equation (14)). The bounds have a simple form and are left in terms of c_N and some Φ functions which correspond to $\mathbb{E}[(\nu_1)_3]$ and $\mathbb{E}[(\nu_1)_2(\nu_2)_2]$, which could easily be computed for a given model. It is clear that under the conditions of convergence to KC, the given TV bound converges to zero, as expected.

Möhle and Sagitov, 2001, "A Classification of Coalescent Processes for Haploid Exchangeable Population Models"

- Constant population size N
- Offspring counts are exchangeable within generations and i.i.d. between generations — implies the genealogical process is a time-homogeneous Markov chain
- Treats full variety of possible limits for exchangeable models (again, I think the family is Ξ-coalescents)
- Presents necessary and sufficient conditions for weak convergence to an appropriate Ξ-coalescent, and the particular Ξ is uniquely determined by moments of the offspring distribution
- A proof of FDD convergence is given. When $c_N \to c \neq 0$ the limiting process is discrete and weak convergence follows immediately from FDDs. When $c_N \to 0$ the limit is continuous and tightness is needed for weak comvergence; the proof is not presented, just a reference to Möhle1999.
- Since the conditions given are necessary and sufficient, and can be applied to any exchangeable (Cannings-type) model, this constitutes a complete characterisation of the limiting genealogies for this class of models.
- Section 6 gives a cute bit of history of coalescent theory

Möhle, 2002, "The Coalescent in Population Models with Time-Inhomogeneous Environment"

Möhle and Sagitov, 2003, "Coalescent Patterns in Exchangeable Diploid Population Models"

Schweinsberg, 2003, "Coalescent Processes Obtained from Supercritical Galton–Watson Processes"

SMC GENEALOGIES

Jacob, Murray, and Rubenthaler, 2015, "Path Storage in the Particle Filter"

Description of ancestries as trunk+crown. Upper bound on storage cost via an approximate multinomial resampling scheme that is independent of weights. Numerical simulations suggesting similar results for stratified and systematic resampling (including an ordering on the schemes?).

Koskela et al., 2018, *Asymptotic genealogies of interacting particle systems with an application to sequential Monte Carlo*

VARIANCE ESTIMATION

Chan and Lai, 2013, "A General Theory of Particle Filters in Hidden Markov Models and some Applications"

Lee and Whiteley, 2018, "Variance Estimation in the Particle Filter"

Olsson and Douc, 2019, "Numerically Stable Online Estimation of Variance in Particle Filters"