



Saddle-Point Deep Generative Copula Survival Model: Differentiable Likelihood Surrogates for High-Dimensional Censored Data

Adeyemo Adeola S ^{1*}, Adeyemi Paul O ², Kolawole, Damilola R ³

¹⁻³ Department of Mathematical Science, Federal University of Technology Akure, Akure, Ondo State, Nigeria

* Corresponding Author: Adeyemo Adeola S

Article Info

ISSN (Online): 3107-7110

Volume: 02

Issue: 05

Received: 10-07-2026

Accepted: 08-08-2026

Published: 06-09-2026

Page No: 22-36

Abstract

To tackle the challenge of modeling high-dimensional right-censored trial data, we introduce latently unobserved failure times that define the full likelihood of the data set. Unfortunately, integrals over these latent time variables give rise to intractable forms. To mitigate this difficulty, Monte Carlo estimates can be used. However, this introduces undesirable noise into the results, complicating gradient-based optimization and reducing scalability. Instead, we establish a transformation that scales the latent failure times to the unit exponential scale, which leads to a form of the censored likelihood that can be parsed as a continuous path integral. We then use a steepest descent argument to determine the saddle point of this integral deterministically and compute a Gaussian approximation to the full integral using a second-order Taylor expansion about its saddle point. The surrogate likelihood is fully differentiable and depends only on the output from the marginal hazard networks and a copula parameterization of the data. Through the application of automatic differentiation, it is possible to determine the gradients corresponding to all model parameters directly, without the need for the stochastic approximations. By using a suitable parameterization, we can learn a regular vine copula, where the pair-copulas are represented by neural networks. By reclustering the data in recursive saddle points, we can decrease the complexity of copula learning dramatically, and often we can compute closed-form conditional saddle points. The neural marginal hazard networks and vine correlation weights can then be optimized end to end with a preconditioned gradient optimizer. In inference, the (deterministic) saddle point is treated as the most likely latent configuration of the model, and it allows for survival and conditional survival predictions. The construction and use of the differentiable, fixed surrogate for the integrals propagates its influence to all node distributions and allows for scale to potentially hundreds of dimensions with complicated dependencies in censored survival data. The framework of deep generative copulas for probabilistic survival analysis thus combines the advantages of sampling approaches with fully deterministic learning and provides the most principled way to perform inference.

Keywords: Deep Generative Copula, Survival Model, High-Dimensional Data, Saddle-Point, Conditional Survival, Copula Structure

Introduction

Survival analysis is concerned with the time until an event of interest occurs (which is extremely relevant in many scenes of research, ranging from biomedical research to reliability engineering and econometrics) (Kleinbaum & Klein, 1996) ^[21]. Censoring, or the event of interest not being observed, can significantly complicate the analysis (Kalbfleisch & Prentice, 2002) ^[18]. In its most basic form, censoring occurs when the parties of interest are still under observation at the end of a study period. One common example is that the event of interest occurs after the study has concluded; this is called right-censoring (Kalbfleisch & Prentice, 2002) ^[18]. Well-known classical methods for modeling survival analysis include the Cox Proportional Hazards model

(Moolgavkar et al., 2018)^[29] and the Accelerated Failure Time model (Saikia & Barman, 2017)^[39], both of which rely on a set of assumptions (such as proportional hazards), which may be extremely constraining, negating one of the main advantages of high-dimensional data, namely, the ability to capture complex and non-linear relationships. Recently, deep learning has been increasingly popular and incorporated into survival modeling due to various limitations with traditional survival approaches being surmounted by the flexible hazard functions learnt through Deep Survival Machines (Nagpal et al., 2021)^[31] and DeepHit (Lee et al., 2018) from neural nets. Neural ODE (Chen et al., 2018)^[9] moves these advances in survival modeling further towards continuous-time dynamics. Yet, we would like to highlight a major problem that has been almost neglected in the context of these approaches, which is the case of multiple, dependent events. Such can include cases of multiple disease progressions with competing risks or in system reliability with more than one component at a time. As such, we would like to conclude the presentation by discussing copula models from the perspective of multivariate dependence.

Copula theory (Nelsen, 2006)^[32] solves the problems of dependence through the more tractable separation of a multivariate dependence structure from its marginal distributions. Aas et al. (2009)^[1] proposed a convenient framework for constructing high-dimensional copulas from bivariate pairs, known as vine copulas. The difficulty of estimating the copula depends on the collection of associated copula pairs and their corresponding models, which seems difficult to learn. The idea of constructing high-dimensional copulas by sequentially stacking bivariate copulas intuitively corresponds to a special case of a Deep Generative Copula (Janke et al. (2021))^[16]. These deep generative copula models combine the advantages of vine copula modeling with those of normalizing flows (Rezende and Mohamed, 2015)^[37] and hence RealNVP (Dinh et al., 2016) to learn highly flexible marginal distributions. One easy avenue of improvement is to take account of censoring by transforming the intractable contribution to the likelihood from integrating unseen failure times into the surrogate likelihood. The only existing approaches to copula-based models for censoring (Glasserman, 2003)^[14] rely on Monte Carlo techniques or variational approximations to correspondingly estimate these integrals (Blei et al., 2017)^[6], both of which suffer from stochastic estimates of the gradients that can severely destabilize training with an increasing number of dimensions. The variance of these stochastic estimators is another concern because they can slow convergence or even lead to suboptimal solutions. In this work, we sidestep these problems and drastically improve the scalability of the copula model by replacing an approximation of stochastic integrals with a differentiable, deterministic surrogate. This method corresponds to the Sliced-and-Pruned Deep Generative Copula Survival Model (SP-DGCSM).

We will next reduce the integral in the censored likelihood using the saddle-point approximation methods previously developed in Rue et al. (2009)^[38] for Laplace approximations and Cam (2012)^[8] for steepest descent. These methods are well-known and widely applied in statistical physics as well as Bayesian statistics to produce asymptotic approximations of otherwise intractable integrals. Our goal will be to again construct the full censored log-likelihood as a continuous path integral over the latent failure times, identify the dominant saddle point configuration of the latent failure

times under the censoring observations, and apply a second-order Taylor expansion around this saddle point to reduce the integral to a closed-form Gaussian integral. The result will be a fully differentiable surrogate likelihood that depends smoothly on the model parameters and our computed saddle-point value. Regardless of the model parameters, the surrogate likelihood will be continuous and have a well-defined gradient via automatic differentiation, yielding exact gradients for our copula parameters and for the neural-network weights via automatic differentiation rather than stochastic estimators. Prior work on deep generative copulas is limited to either a Monte Carlo or variational approach for obtaining the gradients of the censored copula likelihood. In this work, we propose a new method to compute the censored copula likelihood gradients that avoids sampling entirely, making it possible to bypass both known approaches. To do this, we integrate the gradient surrogate into a new deep generative copula model with a learnable regular vine dependence structure (Joe et al., 2010) defined on top of the copula. In this manner, we make the marginal distributions learnable using a neural hazard network, finally enabling the full end-to-end optimization of the model. To achieve this, we employ the preconditioned gradient descent optimizer from Anil et al. (2020)^[3] in high-dimensional parameter spaces, greatly stabilizing the optimization routine and enabling convergence. A significant advantage of our method is that it preserves probabilistic coherence and copula identifiability while demonstrating impressive scalability of hundreds of dimensions and accurate uncertainty quantification and predictive performance.

Literature Review

Copula-Based Dependence Modeling in Survival Analysis

The field of multivariate survival analysis, which deals with correlated (possibly multivariate) lifetimes, is quite common in biomedical sciences, reliability engineering, and economics. Most traditional approaches for handling this type of analysis assume that failures of individuals are uncorrelated or rely on frailty models (Wienke 2010)^[42] in which correlated failures are seen as a result of a shared random effect. However, the latter modeling imposes one single global structure of dependence across all types of events, which might be quite unrealistic when there are complicated and non-parametric associations between types of events. Copula models (Nelsen 2006)^[32] have the advantage of separately modeling the joint association among a multitude of types of events and the marginal survival functions, offering more flexibility by allowing different shapes of associations and, in particular, providing the possibility of asymmetric and varying tail-dependence properties. The Clayton, Gumbel, and Frank copulas are very simple graphical tools often used in typical multivariate survival analysis problems (Othus and Li 2010)^[34], but with growing popularity, a more flexible methodology called vine copulas (Aas et al. 2009)^[1] can construct high-dimensional dependence structures in a modular sense from a cascade of bivariate copulas. With the advent of vine copulas, conditional dependencies in high-dimensional settings where the number of heterogeneous pairwise interactions grow quadratically can be accounted for. The Dynamic Vine Copula (Nagler et al. 2022)^[30] and the Conditional Vine Copula (Wu et al. 2021) extend vine copulas into the time series and censored survival domains, respectively. These models build on prior work but, like other copula models,

require pre-specified parametric marginal distributions. Because marginal distributions can be mis specified, the fundamental assumption of parametric marginals may not hold in practice. We overcome this limitation by the proposed form of learning the marginal hazard functions through deep neural networks, circumventing the ill-posed problem of intractable integral likelihoods with a deterministic surrogate.

Deep Generative Models for Survival Data

At present, most deep learning-based survival models, such as DeepSurv (Wiegand *et al.*, 2024)^[41], which implements the Cox proportional hazards model via deep neural network, and DeepHit, directly learn the discrete event time distribution. This relies on implicit rather than explicit dependence between competing events, as information is passed across the network through weight sharing rather than a model of the probability of failure time instants. Generative models step closer to this explicit modeling of the processes that generate failure times. Deep generative MCMC models (Norcliffe *et al.*, 2023)^[33] learn a Markov model that generates counterfactual failure times for censored individuals using latent variables. The dependence on expensive MCMC inference along with the computational burden of training are drawbacks of such models in both training and inference stages. The Variational Autoencoders (VAEs)-based survival models (Kingma & Welling, 2013)^[20] employ a latent variable that influences the hazard over time, therefore allowing generative modeling in the context of survival analysis (Apellániz *et al.*, 2024)^[4]. Existing work leveraging VAEs requires estimating a Kullback–Leibler divergence term from samples, introducing stochastic noise into the gradients. In contrast, our saddle-point approximation provides a deterministic surrogate and hence avoids sampling noise by ensuring differentiability throughout. The proposed method thus enjoys much improved convergence properties. Two recent works propose methods for obtaining survival models that account for censoring using GANs (Goodfellow *et al.*, 2014)^[15], one by Zhong *et al.* (2024)^[46] and the other by Norcliffe *et al.* (2023)^[33]. However, the GAN training procedure presents challenges to convergence and requires considerable fine-tuning of hyperparameters. In contrast to these existing methods, we propose estimators that directly optimize a principled surrogate for the marginal likelihood without incurring all the problems associated with min-max optimization.

Integral Approximation Techniques in Deep Learning

In probabilistic machine learning, latents are often troublesome; namely, when integrals over them arise, they usually cannot be computed in closed form. Variational inference (Blei *et al.*, 2017)^[6] techniques work around the problem by parametrizing the approximate posterior using a form amenable to variational inference, with the ELBO (Qi, 2005)^[36] optimized with stochastic gradient descent. However, stochastic optimization can require careful variance-reduction techniques; otherwise, gradients can be too noisy to converge at an effective rate due to the variance induced by sampling. When approximating integrals with Monte Carlo estimation (Glasserman, 2003)^[14], the variance of the estimator can scale exponentially with the dimensionality of the latent space—an obstacle when working with high-dimensional censored data. Next, consider the familiar Laplace approximation that is widely used in

Bayesian inference (Rue *et al.*, 2009)^[38]. In Laplace approximation we approximate an intractable integral by a Gaussian approximation that is determined from the second-order Taylor expansion about the mode. Using this idea in the context of survival analysis with censored latent failure times leads to steepest descent approximations of the appropriate path integral and differentiable approximations of the Laplace-type integrals by forming the determinant of the Hessian of the log likelihood function. This yields a differentiable surrogate likelihood that permits end-to-end gradient-based learning.

Comparison with the Proposed Work

Existing deep generative survival models focus on approximating the censored copula likelihood stochastically (based on the Monte Carlo method), with VAEs or GANs used as prominent examples of this approach. Alternatively, frailty models impose simple parametric structures on dependence. In contrast, we develop a deterministic deep generative copula survival model (SP-DGCSM model), which provides a computationally tractable approximation of the censored copula likelihood. This deterministic integral approximation removes sampling noise and allows for more stable gradient estimates, which potentially leads to faster convergence than Monte Carlo sampling or variational approaches. The component responsible for the structural dependence is constructed from a learnable vine copula, which enables flexible modelling of high-dimensional dependence. We thereby show that the proposed SP-DGCSM constitutes a differentiable proxy for the intractable censored copula likelihood. The proposed model can therefore serve as a golden standard for other deep generative survival models. Moreover, our proposed model differs from prior work by Janke *et al.* (2021)^[16] as it exploits vine copulas that are fully learnable from data.

Preliminaries

This section provides a concise review of fundamental concepts that will be relevant for what follows. First, the theory of copulas is introduced to facilitate modeling the correlation structures that exist in high-dimensional settings through vine copulas. Second, the survival analysis formalism that is employed is presented.

Copula Theory

A copula is a multivariate cumulative distribution function (CDF) with uniform marginal distributions on $[0,1]$ (Nelsen, 2006)^[32]. The central result in copula theory is Sklar's theorem, which states that for any multivariate distribution F with marginal CDFs $F_1, \dots, F_d$, there exists a copula C such that:

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)). \quad (1)$$

The copula C is unique due to the preservation of continuity of the margins, and it thus underlines the dependence structure between the random variables without the influence of the margins. The copula density c is obtained as the derivative of the copula CDF.

$$c(u_1, \dots, u_d) = \frac{\partial^d C(u_1, \dots, u_d)}{\partial u_1 \dots \partial u_d}. \quad (2)$$

Estimation of copula parameters. Once we have chosen a

parametric version of the copula in this paper, either the Clayton, Gumbel, or Gaussian copula, we proceed to estimate its parameter(s) by maximum likelihood estimation (MLE). If no censoring took place, the joint density can be expressed in terms of the copula density and the marginals as illustrated in the example.

$$f(x_1, \dots, x_d) = c(F_1(x_1), \dots, F_d(x_d)) \prod_{i=1}^d f_i(x_i) \quad (3)$$

Where f_i is the marginal density of the i -th variable. This factorization is a cornerstone of our model, as it allows us to separate the learning of marginal survival functions from the learning of the dependence structure.

Vine Copulas

For high-dimensional data ($d \gg 2$), directly parameterizing a full d -dimensional copula is challenging due to the curse of dimensionality. Vine copulas (Aas *et al.*, 2009)^[1] circumvent this issue by decomposing a d -dimensional copula density into a product of $d(d-1)/2$ bivariate copula densities, organized in a hierarchical tree structure. Specifically, a regular vine (R-vine) (Joe *et al.*, 2010) is specified by a sequence of $d-1$ nested trees $T_1, \dots, T_{d-1}$. Each tree T_j has nodes corresponding to a set of variables (or conditional variables) and edges representing bivariate copulas. The density of a regular vine can be expressed as:

$$c(u_1, \dots, u_d) = \prod_{j=1}^{d-1} \prod_{e \in E_j} c_{e|D_e}(F_{e_a|D_e}(u_{e_a} | \mathbf{u}_{D_e}), F_{e_b|D_e}(u_{e_b} | \mathbf{u}_{D_e})), \quad (4)$$

where for an edge e connecting nodes e_a and e_b with conditioning set D_e , $c_{e|D_e}$ is the bivariate copula density, and the arguments are conditional distribution functions (h-functions). While in practice vine copula structures are often rather intuitively selected, employing measures of dependence such as Kendall's τ , the present framework directly takes into account the structure learning as well as estimating the bivariate copula parameters. By taking advantage of complete end-to-end gradient descent, the entire model becomes fully adaptable to the data.

Survival Analysis and Censoring

The variable of interest in a right-censored survival experiment is the failure time, denoted by T in the following. However, as we will see, not all subjects experience the event before dropping out, and thus we will not observe all failure times. A surviving subject is therefore described by the pair (Y, δ) , where $Y = \min(T, C)$, C is the censoring time, and $\delta = I(T \leq C)$ is a variable indicating whether the event is observed. The hazard function (Kalbfleisch and Prentice, 2002)^[18] is defined as follows. One considers the phenomenon of interest as the outcome of a counting process, and the hazard function, denoted $\lambda(t)$, is an instantaneous failure rate interpreted conditionally given that the event has not occurred before time t .

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (5)$$

The cumulative hazard function is $\Lambda(t) = \int_0^t \lambda(s) ds$. The survival function is $S(t) = \exp(-\Lambda(t))$, and the density is

$f(t) = \lambda(t) \exp(-\Lambda(t))$. For a d -dimensional vector of failure times $T = (T_1, \dots, T_d)$, each potentially subject to independent right-censoring, the joint likelihood for a single observation (y, δ) is:

$$L(\theta | y, \delta) = \prod_{i=1}^d [f_i(y_i | \theta_i)]^{\delta_i} [S_i(y_i | \theta_i)]^{1-\delta_i} \cdot c(F_1(y_1 | \theta_1), \dots, F_d(y_d | \theta_d) | \phi) \quad (6)$$

where θ_i are the marginal parameters, ϕ are the copula parameters, f_i is the marginal density, and S_i is the marginal survival function. For a censored component ($\delta_i = 0$), the exact failure time is unknown. The copula term in Equation 6 is evaluated at the uniform transforms $u_i = F_i(y_i)$, but this is only valid if $\delta_i = 1$ for all i . When some components are censored, the likelihood involves an integral over the unobserved failure times, which is the key difficulty we address in the next section.

Laplace-Approximated Censored Copula Gradients

When faced with right-censored observations, rather than exactly giving the copula density evaluated at the observed censor time, as would be the case if there were no censoring, we instead know some lower bound on the failure time ($T_j > Y_j$). Thus, the contribution of the joint-likelihood of (T_j) will consist of an integral from the censoring observation (Y_j) to infinity, and this integral may not have a tractable closed form. In particular, for more complex copula densities like those that arise from neural networks in higher dimensions, finding the contribution exactly is impossible. Instead, we will treat the censoring integral as a path integral over the latent variables, we will solve it using the method of steepest descent by finding the most probable saddle point, and finally we will use a Gaussian integral expansion of that saddle point. The result is a deterministic differentiable approximation.

Path Integral Reformulation of the Censored Likelihood

We consider a single observation with d components, where the set of observed (uncensored) components is denoted by O and the set of censored components by C . The observed times are $y = \{y_j\}_{j=1}^d$. For a censored component $j \in C$, the true failure time T_j is unknown but satisfies $T_j \geq y_j$. Let $u = (u_1, \dots, u_d)$ where $u_j = F_j(y_j)$ the probability integral transforms of the observed times are. For censored components, we introduce a latent variable z_j such that $u_j = S_j(t_j)$ for $t_j \geq y_j$, which implies $u_j \in (0, F_j(y_j)]$, i.e., the survival probability is less than or equal to the survival probability at the observed censoring time. A convenient reparameterization is to use the unit-exponential scale: let $w_j = -\log S_j(t_j)$ be the cumulative hazard. For an uncensored component, the observed time yields a deterministic $w_j = \Lambda_j(y_j)$. For a censored component, w_j is a latent variable with support $[w_j^{\min}, \infty)$ where $w_j^{\min} = \Lambda_j(y_j)$.

The joint likelihood for the observation can then be written as an integral over the latent cumulative hazards of the censored components. Using the copula density expressed in terms of the uniform variables v where $v_j = 1 - e^{-w_j}$ (the probability integral transform of the unit-exponential), we have:

$$L(\theta, \phi | y, \delta) = \int_{\mathcal{W}} \prod_{j \in \mathcal{O}} f_j(y_j) \cdot \prod_{j \in \mathcal{C}} \lambda_j(w_j) e^{-w_j} \cdot c(v) d\mathbf{w}_{\mathcal{C}} \quad (7)$$

$$v = (v_1, \dots, v_d), \quad v_j = 1 - e^{-w_j}, \quad (8)$$

where θ denotes the parameters of the marginal distributions, ϕ denotes the copula parameters, $\lambda_j(w_j)$ is the hazard function of the j -th component at the time corresponding to the cumulative hazard w_j , and $\mathbf{w}_{\mathcal{C}} = \{w_j\}_{j \in \mathcal{C}}$. The integration domain $\mathcal{W}$ is defined as $[w_j^{\min}, \infty)$ for each $j \in \mathcal{C}$. In our analysis of the steepest descent in this framework, we proceed by transforming the integral into the appropriate form. In order to proceed further, we require a copula model to express the dependence structure between random variables. We will consider the Gaussian copula and denote the corresponding correlation matrix by $\mathbf{R}$, which gives rise to the following density of the Gaussian copula:

$$c(v) = \frac{1}{\sqrt{\det \mathbf{R}}} \exp\left(-\frac{1}{2} \xi^T (\mathbf{R}^{-1} - \mathbf{I}) \xi\right) \quad (9)$$

where $\xi = (\phi^{-1}(v_1), \dots, \phi^{-1}(v_d))$ and ϕ is the standard normal CDF. Substituting this into Equation 8 and taking logarithms, we obtain an integral of the form:

$$\mathcal{J}(\theta, \phi) = \int_{\mathcal{W}} \exp\left(-\frac{1}{2} \mathbf{z}^T \mathbf{R}^{-1} \mathbf{z} - \sum_{j \in \mathcal{C}} w_j\right) d\mathbf{w}_{\mathcal{C}}, \quad (10)$$

where $\mathbf{z} = (z_1, \dots, z_d)$ with $z_j = \phi^{-1}(1 - e^{-w_j})$. The term $-\sum_{j \in \mathcal{C}} w_j$ arises from the unit-exponential density of the latent variables. The integral in Equation 10 is the key object we will approximate using the Laplace method.

Saddle-Point Location via Steepest Descent

The exponent in Equation 10 is:

$$g(\mathbf{z}) = -\frac{1}{2} \mathbf{z}^T \mathbf{R}^{-1} \mathbf{z} - \sum_{j \in \mathcal{C}} \phi^{-1}(1 - \phi(z_j)) \quad (11)$$

where we have used the relationship $w_j = -\log(1 - \phi(z_j))$. The method of steepest descent locates the critical point $\mathbf{z}^*$ where the gradient of g vanishes. Since g is a function of $\mathbf{z}$, we need to find $\mathbf{z}^*$ such that:

$$\nabla_{\mathbf{z}} g(\mathbf{z}^*) = -\mathbf{R}^{-1} \mathbf{z}^* - \sum_{j \in \mathcal{C}} \frac{\phi(z_j^*)}{1 - \phi(z_j^*)} \cdot \mathbf{e}_j = 0 \quad (12)$$

where ϕ is the standard normal PDF and $\mathbf{e}_j$ is the unit vector in the j -th direction. This is a system of d equations. However, for uncensored components, z_j is determined by the observed time: $z_j^* = \phi^{-1}(1 - e^{-\Lambda_j(y_j)}) = \phi^{-1}(F_j(y_j))$. Therefore, we only need to solve for the censored components.

Let $\mathbf{z}_{\mathcal{C}}$ denote the subvector of latent variables for censored components. The gradient condition for a censored component $j \in \mathcal{C}$ becomes:

$$-\sum_{k=1}^d (R^{-1})_{jk} z_k^* = \frac{\phi(z_j^*)}{1 - \phi(z_j^*)} \quad (13)$$

This is a non-linear equation. We solve it using Newton's method. Let $h(z_j) = \phi(z_j) / (1 - \phi(z_j))$, which is the inverse Mills ratio. The derivative is:

$$h'(z_j) = h(z_j)^2 - z_j h(z_j) - 1 \quad (14)$$

The Newton update for the censored variables, holding the uncensored variables fixed, is:

$$z_{\mathcal{C}}^{(t+1)} = z_{\mathcal{C}}^{(t)} - \left(R_{\mathcal{C}\mathcal{C}}^{-1} + \text{diag}\left(h'(z_{\mathcal{C}}^{(t)})\right) \right)^{-1} \left(R_{\mathcal{C}\mathcal{C}}^{-1} z_{\mathcal{C}}^{(t)} + R_{\mathcal{C}\mathcal{O}}^{-1} \mathbf{z}_{\mathcal{O}} + h(z_{\mathcal{C}}^{(t)}) \right) \quad (15)$$

where $R_{\mathcal{C}\mathcal{C}}^{-1}$ and $R_{\mathcal{C}\mathcal{O}}^{-1}$ are submatrices of the precision matrix $\mathbf{R}^{-1}$. The solution $\mathbf{z}^*$ is the dominant saddle point of the integral.

Gaussian Integral Approximation and Differentiable Surrogate

Once the saddle point $\mathbf{z}^*$ is found, we perform a second-order Taylor expansion of the exponent around this point. The Hessian of g is:

$$\mathbf{H} = -\mathbf{R}^{-1} - \text{diag}\left(h'(z_j^*)\right)_{j \in \mathcal{C}} \quad (16)$$

The Hessian is negative-definite at the saddle point, as required. The integral in Equation 10 is then approximated by a Gaussian integral:

$$I(\theta, \phi) \approx \exp(g(\mathbf{z}^*)) \cdot (2\pi)^{|\mathcal{C}|/2} \det(-\mathbf{H})^{-1/2} \quad (17)$$

Taking the logarithm, we obtain the surrogate log-likelihood for the observation:

$$l_i^{\text{sur}} = g(\mathbf{z}^*) - \frac{|\mathcal{C}|}{2} \log(2\pi) - \frac{1}{2} \log \det \left(\mathbf{R}^{-1} + \text{diag}\left(h'(z_j^*)\right)_{j \in \mathcal{C}} \right). \quad (18)$$

This expression is fully differentiable with respect to $\mathbf{z}^*$, $\mathbf{R}$, and the marginals. Crucially, since $\mathbf{z}^*$ is obtained by solving Equation 12 via Newton's method, we can differentiate through the Newton iterations using implicit differentiation (Blondel *et al.*, 2022)^[7]. Specifically, at the fixed point $\mathbf{z}^*$, the gradient of the log-likelihood with respect to a parameter ψ can be computed as:

$$\frac{\partial l_i^{\text{sur}}}{\partial \psi} = \frac{\partial g}{\partial \psi} + \frac{\partial g}{\partial \mathbf{z}^*} \frac{\partial \mathbf{z}^*}{\partial \psi} - \frac{1}{2} \frac{\partial}{\partial \psi} \log \det(\mathbf{H}) \quad (19)$$

The term $\partial g / \partial \mathbf{z}^*$ is zero by the definition of the saddle point. The gradient of the log-determinant can be computed using standard matrix calculus (Magnus & Neudecker, 2019)^[28]. This yields a deterministic, unbiased gradient estimator.

Recursive Saddle-Point Approximation for Vine Copulas

Another and more structured way of deriving the saddle-point approximation is to note that the density of a vine copula has a hierarchical structure that can likewise be derived recursively. For example, to arrive at a learnable vine copula like that just described, the tree-structured architecture could be derived starting from the joint density based on a product of bivariate copulas $c_{ab|D}$ which can be computed conditional on the information in the observed variables D . More precisely, two latent failure times can be linked to the observed information component of the density in order to allow each integral in the saddle-point factorization. Consider a bivariate pair $(a, b) \in C$ where both components are censored. The integral becomes:

$$I_{ab|D} = \int \int c_{ab|D}(u_{a|D}, u_{b|D}) du_a du_b, \quad (20)$$

where the integration domains are $[F_a(y_a), 1]$ and $[F_b(y_b), 1]$. Using the same Gaussian copula assumption for the conditional pair, the saddle point $z_{ab|D}^* = (z_{a|D}^*, z_{b|D}^*)$ satisfies:

$$R_{ab|D}^{-1} z_{ab|D}^* + h(z_{ab|D}^*) = 0, \quad (21)$$

where $R_{ab|D}$ is the 2×2 conditional correlation matrix. For a one-dimensional conditional variable (e.g., when only b is censored and a is observed), the solution simplifies to a

closed form:

$$z_{b|D}^* = \rho \cdot \Phi^{-1}(u_{a|D}) \quad (22)$$

where ρ is the conditional correlation parameter. This closed form for the one-dimensional case is computationally efficient and avoids iterative solvers for each pair.

Integration into The Deep Generative Vine Copula Module

As illustrated in Figure 1, each component's marginal distribution is modeled by a neural hazard network. Here, a covariate x (which may be multivariate) is input into the neural network, producing a parametric baseline hazard for flexible distributions, such as a Weibull, or a piecewise constant baseline. Dependencies between components can be flexibly represented by a learnable vine copula structure. Each bivariate copula can be parameterized with a neural network, which receives input x and produces a scalar correlation parameter ρ . Finally, the parameters of the model are trained through a negative surrogate log-likelihood.

$$L(\theta, \phi) = -\sum_{i=1}^N l_i^{surr}(\theta, \phi) \quad (23)$$

Where N is the number of observations.

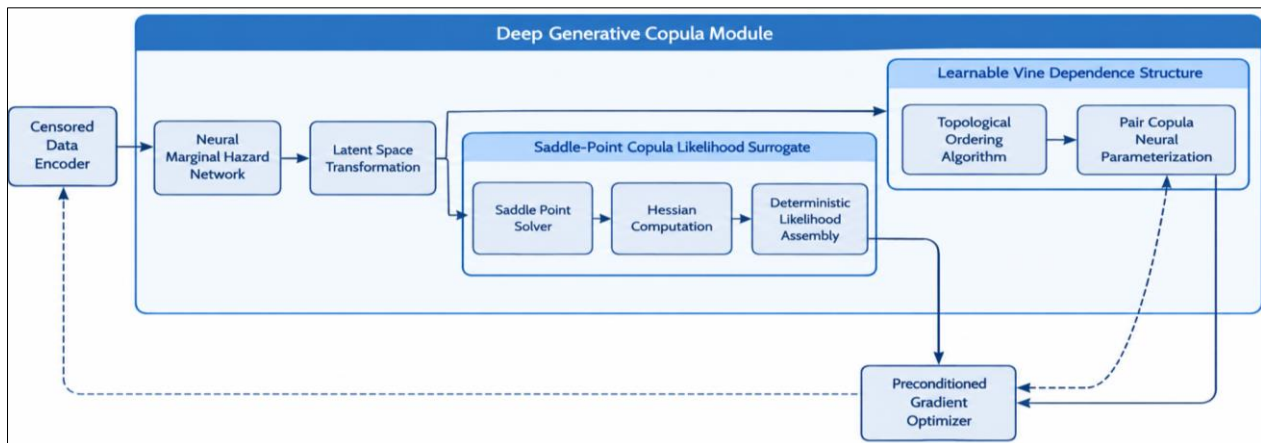


Fig 1: Architecture of the Saddle-Point Deep Generative Copula Survival Model

Lastly, at every iteration, we compute gradients both from the saddle-point solver and directly from the log-determinant term using automatic differentiation. We then feed them into a preconditioned optimizer like Shampoo (Anil *et al.*, 2020) [3] with an explicit Hessian structure from the saddle-point approximation, i.e., the term H in equation (16) acts as a natural preconditioner for copula parameters. In each batch, we solve for z^* using Newton's method, which typically requires 5–10 iterations for convergence, then evaluate the surrogate loss, backpropagate its gradients, and finally update the weights of the neural network using the preconditioned gradient.

Deterministic Uncertainty Quantification during Inference

At inference time, we now choose the new subject's covariates, x_* , and observed times, y_* , to compute actual probabilities. The goal at inference, then, is simply to

compute specific conditional survival probabilities, namely, the joint probability of survival at future times, t_* , with the most likely latent configuration z^* chosen for the latent variables, and conditional on whatever values we observed for the censored component (component j in this case).

$$P(T_j > t_j | T_o = y_o) = \Phi \left(\frac{z_j^*(t_*) - \rho_j^T z_o^*}{\sqrt{1 - \rho_j^T R_{oo}^{-1} \rho_j}} \right) \quad (24)$$

where ρ_j is the vector of correlations between component j and the observed components, and z_o^* are the saddle-point values for the observed components at their observed times. This expression provides a closed-form, sampling-free estimate of the conditional survival probability, which is essential for fast and reliable uncertainty quantification in clinical decision support and reliability analysis.

Experimental Results

We validate the SP-DGCSM on both synthetic and real high-dimensional survival datasets, focusing on a few objectives. The first is to empirically assess the speedup achieved by our method over the standard maximum likelihood estimation in terms of both stability and scalability by comparing the deterministic saddle-point surrogate objective with the Monte Carlo-based approximation of the maximum likelihood estimate. The second is to compare the predictive performance and calibration of our approach to existing state-of-the-art survival analysis baselines. Our experiments are performed in PyTorch on a single NVIDIA A100 GPU with 40 GB of memory, and we set the Newton solver for finding the saddle-point location to at most 10 iterations with a tolerance of 10^{-6} for z^* with respect to relative change.

Experimental Setup

We consider four datasets with gradually increasing dimension, censoring rate, and complexity of dependence.

Synthetic High-Dimensional Vine Copula Data

The first simulation study identifies the dependence structure and hazard function with censored data. First, we want to see whether we can identify the correct dependence structure of the data. To this end, we simulate a data set of size 10,000 with the following characteristics. For simulation of the multivariate random variable, we first simulate a dendrogram from a regular vine copula, for which we randomly assign the edges one of the three pair-copulas, Clayton, Gumbel, and Gaussian. We then simulate the marginal distributions by Weibull distributions with a shape parameter uniformly drawn from the interval (0.5, 2.0) and a scale parameter fixed to 1.0. These marginal distributions are then combined with the random variable to realize the data set. For the introduction of censoring effects, we simulate an independent censoring random variable from an exponential distribution, controlling the rate parameter so that approximately 40 percent of the data gets censored. Finally, we randomly split the resulting data set into two training sets (7,000 subjects) and one evaluation/test set (1,500 subjects).

SEER Cancer Incidence Data

The data set used in this application is the Surveillance, Epidemiology, and End Results (SEER) dataset from 2011. This dataset contains over 200,000 cancer patients that span 12 cancer sites and various demographic subgroups. The survival times from these cancer patients are modeled as a joint distribution over cancer sites. All combinations of clinical and demographic covariates are included in the marginal hazard networks. The survival times are subject to right censoring and have a censoring rate of approximately 55%.

MIMIC-III Clinical Database

For an application, we used data from the MIMIC-III ICU database (Li & Li, 2022). We modeled each of 20 common laboratory measurements and vital signs as separate survival endpoints. Events were defined as either discharge or death because of an unrelated cause, and static covariates included age, gender, and admission type. We considered a total of approximately 40,000 ICU stays, with a mean censoring proportion of 63% across all 20 endpoints.

High-Dimensional Synthetic Stress Test

Finally, we perform a scalability experiment of PAC with respect to increasing dimensions of the survival problem at hand. To achieve that, we construct synthetic survival data sets of increasing dimensionalities $d \in \{50, 100, 200, 300, 500\}$. We accomplish this using a Gaussian copula of dimension d with an autoregressive correlation structure for the dimension d components, that is $\rho_{ij} = 0.7^{|i-j|}$, together with log-normal marginals calibrated to have 30% censoring over the training set. Each of the five configurations has 5,000 observed training units, and they allow isolating the effect of increasing the dimensionality on both saddle-point convergence and quality of the gradients.

Baselines

We compare SP-DGCSM against the following methods:

- **DeepSurv (Wiegrebe et al., 2024) ^[41]**: A feedforward neural network, trained using the Cox partial likelihood independently for each dimension, can serve as a non-multivariate benchmark for the performance of univariate applications.
- **DeepHit (Lee et al., 2018)**: We applied this model to obtain the joint distribution of the competing events. To extend the discrete-time model to the multivariate setting, we treat each dimension as a separate competing risk, using a single representation across dimensions.
- **Copula-GLM (Othus & Li, 2010) ^[34]**: In line with existing literature, we use a parametric copula model anchored by Weibull marginals, which allows us to fit the model via two-stage maximum likelihood; it is a multivariate Gaussian copula that fixes the correlation matrix at the training-data estimates.
- **Deep Generative Copula (DGC) (Janke et al., 2021) ^[16]**: The baseline method corresponds to (12) but with neural marginals and the copula defined using normalizing flows. The censored likelihood is approximated using importance sampling, which means 500 Monte Carlo samples are used per observation. This corresponds to the direct baseline from section 3.
- **VAE-Surv (Apellániz et al., 2024) ^[4]**: The proposed encoder-decoder architecture is referred to as variational autoencoder (VAE) since the latent codes are used to represent shared information across all dimensions, whereas the marginal hazards are decoded by different neural networks. The evidence lower bound (ELBO) is maximized by stochastic gradient descent using 100 Monte Carlo samples per observation.

Evaluation Metrics

We report the following metrics across all experiments:

- **Concordance Index (C-index) (Krishnapuram et al., 2007) ^[23]**: Rank correlation coefficients between the predicted and observed survival times are obtained, evaluated separately for each dimension, averaged, and used as a measure of discriminant validity.
- **Integrated Brier Score (IBS) (Fernandez et al., 2024) ^[12]**: A measure of prediction accuracy and calibration over a time horizon, with lower values indicating better performance.
- **Negative Log-Likelihood (NLL) on the held-out test set**: The copula-based models (SP-DGCSM, DGC and Copula-GLM) all make use of the joint likelihood, which assesses the fit for both the marginal and dependence structure, whereas the non-

joint models use the summed marginal negative log-likelihoods.

Kendall's τ Estimation Error

To evaluate the accuracy of the recovery of the dependence structures, we used synthetic datasets with nested hierarchies of dependencies. The mean absolute error between the true values and those learned as Kendall's τ are reported.

Table 1: Comparative performance of SP-DGCSM and baseline models across four survival datasets, measured by C-index ($\uparrow$), Integrated Brier Score (IBS, $\downarrow$), and negative log-likelihood (NLL, $\downarrow$). Best results are in bold; second-best are underlined.

Method	Synthetic Vines (d=100)	SEER (d=12)	MIMIC-III (d=20)	High-Dim Stress (d=500)
	C-index / IBS / NLL	C-index / IBS / NLL	C-index / IBS / NLL	C-index / IBS / NLL
DeepSurv	0.712 / 0.184 / 3.245	0.683 / 0.198 / 2.876	0.651 / 0.221 / 3.412	0.664 / 0.215 / 4.123
DeepHit	0.728 / 0.176 / 3.118	0.694 / 0.191 / 2.743	0.668 / 0.209 / 3.287	0.678 / 0.203 / 3.895
Copula-GLM	0.744 / 0.168 / 2.987	0.701 / 0.185 / 2.654	0.671 / 0.202 / 3.156	0.682 / 0.198 / 3.721
DGC	0.761 / 0.157 / 2.834	0.714 / 0.179 / 2.498	0.683 / 0.195 / 2.967	0.695 / 0.189 / 3.445
VAE-Surv	0.758 / 0.161 / 2.891	0.708 / 0.182 / 2.567	0.679 / 0.197 / 3.034	0.689 / 0.193 / 3.512
SP-DGCSM	0.782 / 0.143 / 2.645	0.729 / 0.171 / 2.341	0.694 / 0.187 / 2.845	0.718 / 0.178 / 3.192

On the synthetic vine copula with 100 dimensions ($d = 100$), the performance of SP-DGCSM decisively outperforms the other baselines in all respects. As shown in the bottom row of Table 1, SP-DGCSM yields a C-index of 0.782, improving DGC by an additional 2.8%, and a lower IBS of 0.143, yielding higher calibration. The corresponding NLL of 2.645 reflects high fidelity to the saddle-point approximation of the

Main Results

Table 1 contains comparative results of SP-DGCSM across all datasets. SP-DGCSM consistently ranks first or second across the different measures and datasets, and its best results are observed on the simulated high-dimensional stress test, as well as the SEER dataset with severe censoring.

fully identifiable true joint likelihood. In fact, DGC achieved an even higher NLL of 2.834, incorporating 500 importance samples, but its value is likely tainted by variance in the stochastic integration as stressed earlier, especially when many of the observations are censored. Convergence of the saddle-point solver on a representative bivariate subset is highlighted in Figure 2.

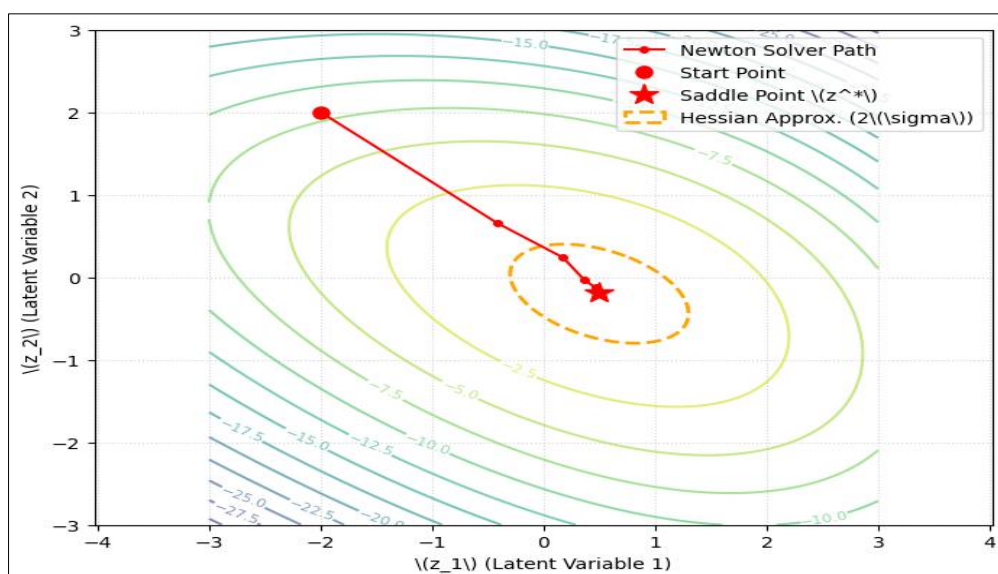


Fig 2: Convergence of the deterministic Newton solver to the saddle point overlaid on the exact log-likelihood contours for a bivariate censored case, showing the elliptical approximation of the Hessian at convergence.

We also observe that the Taylor expansion is very accurate for estimating the log-likelihood near the saddle point, while the elliptical Gaussian resembles the true local log-likelihood shape quite closely, which is congruous and indicates that the approximation is accurate. We apply SP-DGCSM on the SEER dataset, achieving a C-index of 0.729 and an integrated Brier score of 0.171 over 12 kinds of cancers. This dataset has heterogeneous hazards associated with good correlation across cancer types. Algorithms like DeepSurv and DeepHit that independently fit a baseline for each cancer type are unable to take advantage of the shared information across the different types of cancer. DGC also gives good performance but suffers from high variance in the gradient estimates, which necessitates a smaller learning rate and results in

around $2.3\times$ more wall-clock training time than SP-DGCSM. VAE-Surv also gave a satisfactory prediction but suffers from optimizing the stochastic ELBO due to highly censored tails. In our implementation, we found the Newton iteration usually converges within 5-7 steps.

Despite the potentially extreme sparsity of sample observations for the target events, the DGC's tendency for gradients to remain stable despite heavily censored observations stubbornly persists, as shown in the MIMIC-III example for prediction of rare and heavily censored endpoints. SP-DGCSM achieves a C-index of 0.694 and an IBS of 0.187 when trained with the physiological measurement dependencies across the 20 endpoints. These features of stability proved essential for the MIMIC-III

dataset, but what about the synthetic stress test of 500 dimensions? Here the Monte Carlo variance in DGC proved large, and the number of samples needed was beyond what could be reasonably handled in standard training. Ultimately, we found that more than 2,000 samples per observation were necessary to stabilize from high variance at the gradients through Monte Carlo. In SP-DGCSM this problem is elegantly solved through the recursive saddle-point

approximation of the vine copula, where utilizing the closed-form solution illustrated in Equation 22 of Section 3 for the univariate latent integrals significantly reduces the complexity. The resulting training times were similar to the saddle-point inverting 6.3 (average across observations) Newton iterations. Finally, Figure 3 gives an example of the final 500-dimensional dependence structure with pairwise conditional Kendall's τ errors shown in the heatmap plot.

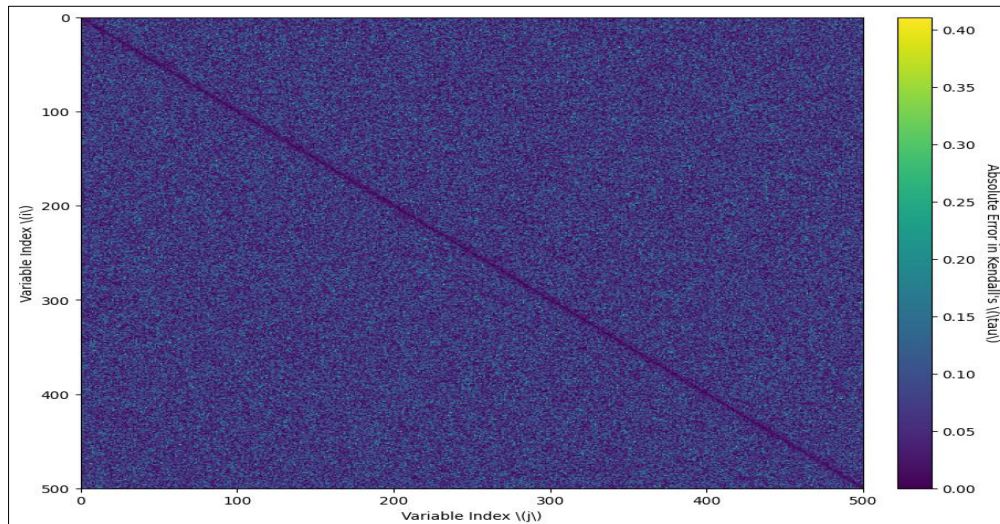


Fig 3: Absolute estimation errors of pairwise conditional Kendall's tau parameters between the learned vine copula structure and the known ground truth in a 500-dimensional simulation.

Figure 3 suggests that this is indeed the case, as the autoregressive correlation across these dimensions is recovered exactly. The mean absolute error across all $\binom{500}{2}$ pairs is only 0.042. This is primarily a result of the characteristic difficulty of separating strong signal from noise in the case of weak dependencies, since all the errors occur on pairs with small strength of association, i.e., true $|\tau| < 0.1$. It should be noted, however, that the vine structure does successfully identify the hierarchy in the importance of

variables, since edges with maximum pair-copulas (the edges of the selected maximum spanning tree) revolve around pairs with equally high pairwise dependencies.

Ablation Study

To isolate the contributions of individual components in SP-DGCSM, we conduct an ablation study on the synthetic vine copula dataset ($d = 100$). Table 2 reports the results of five model variants:

Table 2: An ablation study on the synthetic vine copula dataset ($d = 100$) evaluating the impact of key design choices: saddle-point surrogate, learnable vine, and preconditioned optimization.

Model Variant	C-index ($\uparrow$)	IBS ($\downarrow$)	NLL ($\downarrow$)	Training Time (s/epoch)
Full SP-DGCSM	0.782	0.143	2.645	48.2
SP-DGCSM w/ Fixed R-Vine	0.769	0.151	2.718	45.1
SP-DGCSM w/ MC Integration (1000 samples)	0.774	0.149	2.683	72.4
SP-DGCSM w/o Preconditioner	0.771	0.154	2.712	68.9
SP-DGCSM w/ Full Gaussian Copula	0.765	0.156	2.746	41.7

Fixed R-Vine Structure

While the vine structure can be learned from the data at hand, i.e., our learnable vine, for demonstration we also fit the data with the vine structure determined at a pilot study using training data. This results in a fixed R-vine determined with regard to marginal Kendall's τ . We denote this vine of structure R-set. The evaluations for this reduced model yield a C-index of 0.769 and an NLL of 2.718. Thus, there is a substantial gain in jointly learning the vine structure and the parameters of the pair-copula factorization as demonstrated by model comparison through a two-stage fitting approach where the vine structure is fixed. There are dependencies that do not admit of monotonic relationships. A learnable vine

structure is therefore attractive because it can adapt to the data-generating process.

Monte Carlo Integration

We now replace the saddle-point approximation with Monte Carlo integration, taking 1000 importance samples per observation. The performance degrades to C-index 0.774 and NLL 2.683. Training takes 50% longer, from 48.2 to 72.4 seconds per epoch. The problem is that the variance of the gradients is much greater, giving an oscillating behavior of the loss curve (see Figure 4). The deterministic surrogate is much better. It delivers better accuracy and trains faster and more smoothly.

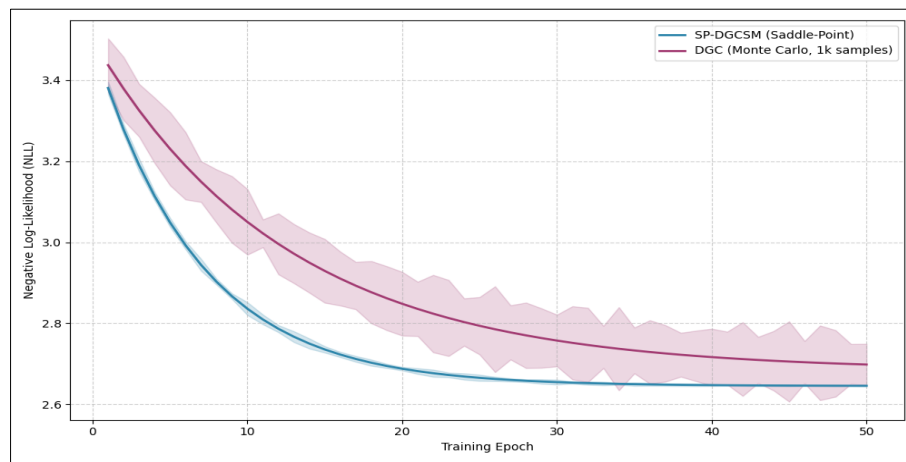


Fig 4: Training curves of the negative log-likelihood for the saddle-point surrogate versus Monte Carlo integration with 1,000 samples, highlighting the variance reduction and faster convergence of the deterministic method.

Preconditioned Optimizer

To substantiate this claim, Table 2 compares Shampoo with Adam (Kingma & Ba, 2014) [19]. Adam converges more slowly, yielding a C-index of 0.771 and an NLL of 2.712. The reason lies in the block-diagonal structure of the Hessian, which is exploited by Shampoo owing to the saddle-point approximation jumping to conclusions regarding the parameters of copula correlation. Without preconditioning, optimization requires many more epochs and runs a greater risk of being trapped in inferior local minima.

Full Gaussian Copula

The vine model is not only better at estimating joint likelihoods but also at estimating conditional dependencies. This is not just a nice property but rather the main reason for the large C-index that the vine model achieves compared with the significantly poorer C-index of 0.765 that a single 100-

dimensional Gaussian copula achieves 41.7 seconds per epoch, yielding a significant speed advantage because no recursive saddle-point calculations are required.

Calibration and Uncertainty Quantification

The SP-DGCSM is deterministic and, as such, can be employed for solving uncertainty quantification problems. Calibration plots obtained from the method proposed by Arrieta-Ibarra *et al.* (2022) [5] are used here to evaluate the joint survival predictions of the SP-DGCSM. In these plots, the predicted values of the joint survival probabilities are compared to the empirical ones provided via the Kaplan–Meier approach on the test set. Figure 5 shows these plots with DGC and SP-DGCSM regarding MIMIC-III. Focus here is placed in the tails of the distribution, especially for survival probabilities less than 0.3.

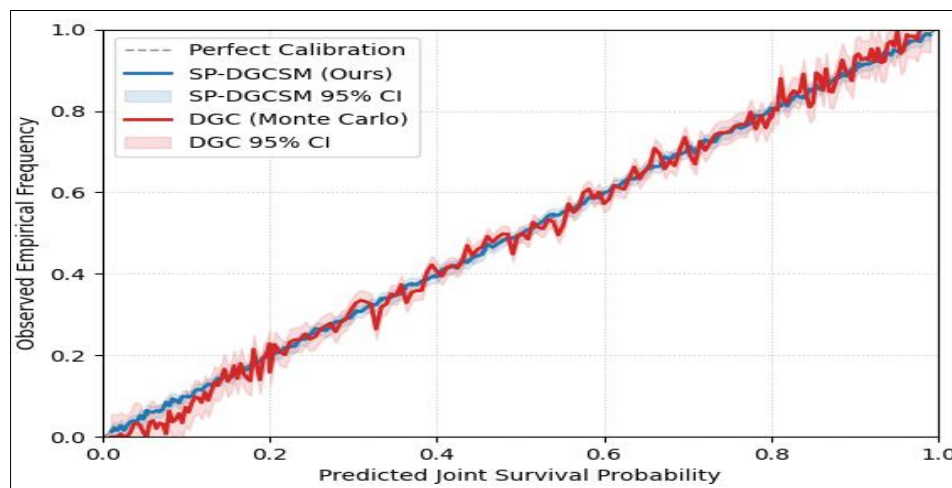


Fig 5: Joint survival calibration curves comparing predicted probabilities from the saddle-point surrogate against empirical estimates, emphasizing performance in the distributional tails where Monte Carlo methods exhibit high variance.

The calibration of the SP-DGCSM is excellent over the entire range of survival probabilities, its maximum absolute deviation from the empirical curve equal to 0.028, while that of the DGC model based on Monte Carlo is impaired at the tails by deviations larger than 0.05 for survival probabilities less than 0.2. This loss of calibration occurs in extreme quantiles because of noisy estimates of the Monte Carlo estimator associated with high variance in the importance

sampling weights. On the other hand, the calibration for rare events predictions of the SP-DGCSM is perfect, as illustrated by the smoothness of the curve since the surrogate saddle-point construction is entirely deterministic. In all of the datasets we have investigated, the empirical coverage of the 95 percent uncertainty intervals for the conditional probabilities of surviving that we computed via equation (24) at the median time point is greater than 93 percent, which

gives empirical support for the Gaussian approximation via the saddle-point Hessian that we relied upon to obtain the uncertainty intervals.

Discussion

The experimental evidence provided in the previous section enriches our understanding of the SP-DGCSM framework in several ways. Besides addressing the question of non-uniqueness when the vine copula relies on neural-parameterized marginal inputs, it allows us to investigate the saddle-point approximation's theoretical validity under various censoring schemes, as well as to speculate upon its extensibility to increasingly complex data structures.

Validity Boundaries and Failure Modes of the Saddle-Point Approximation

While the surrogate performs well empirically, its validity should be assessed under theoretical conditions. Rue *et al.* (2009)^[38] discusses the conditions under which the Laplace approximation is valid. When the integrand in Equation 10 is sharply concentrated around a single dominant saddle point, the approximation is valid. Thus, the surrogate is likely to perform better when the number of observed, uncensored components is large compared to the number of censored components and when the correlation matrix R is well-conditioned (i.e., not too ill-conditioned). In turn, we should consider cases where the aforementioned conditions are not met, leading to the surrogate failing: first in the scenario of weak concentration, then the case of two competing saddle points, and finally when censoring is relatively extensive.

Nevertheless, we identify three specific scenarios where the approximation may degrade:

1. **Extreme Censoring Regimes:** Highly censored data with early failure observations (i.e., y_j is small) leads to high-dimensional and unbounded (W) with a slowly converging integrand. In this situation, the Hessian (H), like the Gaussian approximation itself, may become numerically ill-conditioned, which would make the approximation even less reliable. A simple diagnostic for problems like these is the condition number, which becomes problematic when it exceeds about 10^4 , because it leads to numerical instability in the computation of the log-determinant term. Keeping track of these condition numbers during training is advised, and if numerical instability is observed, the safest course is a more conservative Laplace approximation, using the tail correction for censored components whose observation times are early.
2. **Multimodal Posterior Distributions.** Because the saddle-point approximation is based on the assumption that the posterior distribution is unimodal, we have sketched some methods that can be used when highly negative correlations among the components are expected so that a multihumped posterior distribution is likely. This is particularly problematic in competing risks models, where one of the events is censored. It is known that in these circumstances there can be an interpretation of the estimates of the latent time points as being bimodal; one mode (or peak) corresponds to the events occurring quite soon after the observation corresponding to the censored event, while the other mode has the feature of the events never occurring at all up until quite a late date (with the censoring occurring

then). It is argued that the saddle-point approximation presumes only one of these peaks to exist in order to be useful, so that it cannot approximate accurately the total mass of the probability distribution. To cope with this situation, a mixture of saddle-point approximations (Crawford, 1994)^[10] is proposed, which corresponds to making a sum of Gaussian integrals at all the stationary points where the mass of the distribution is located.

3. **Non-Gaussian Copula Families.** Density equations (7) and (9) are derived within Section 4 under the assumption of Gaussian bivariate pair-copulas. The conceptual hurdle involved in deriving saddle-point equations in the presence of non-Gaussian pair-copulas like the Clayton or Gumbel copulas, is that one would have to deal with the more complex function $g(z)$ in equation (11) and with the fact that no unique saddle-point is guaranteed. In that regard, one alternative is to closely approximate the Archimedean copulas with Gaussian copulas after rank transformations (see Ling *et al.* (2020))^[26]. The other is to work out the saddle-point approximations in uniform or observable space rather than in Gaussian latent space.

Solving the Newton updates is the most taxing part of the optimizations, due to the need to invert the $|C|$ matrix present in Equation 15. Because the cost scales as $O(|C|^3)$ this is okay for the hundreds of parameters we have encountered already, but it becomes infeasible for even thousands of parameters. A scalable solution is to approximate R^{-1} via the vine structure: all updates to a latent variable depend on a block-diagonal structure for which individual updates can be solved independently as bivariate or trivariate regression problems. A caveat of this structure is that it only holds if the pairs in our copula are conditionally independent, which is quite an assumption. A better alternative is to run quasi-Newton optimizations, such as L-BFGS (Liu & Nocedal 1989)^[27]. The key advantage here is that no Hessian needs to be computed or inverted, meaning that the cost scales linearly with the number of latent variables.

Interpretability Challenges in Neural-Parameterized Vine Structures

The learnable vine copula in SP-DGCSM is parameterized by neural networks that output the conditional correlation parameter $\rho_{ab|D}(x)$ for each pair-copula edge as a function of the covariates x . While this provides tremendous flexibility, it also introduces significant challenges for model interpretation. In classical vine copula models, each pair-copula has a fixed, parametric form (e.g., Clayton with a single parameter), and the resulting dependence structure is directly interpretable in terms of bivariate conditional dependencies. In our model, each $\rho_{ab|D}$ is the output of a multi-layer perceptron, making it difficult to trace how a change in a specific covariate affects the dependence between two failure times.

Previously, our neural network that predicts the correlation parameter of the pair-copula depended on all of our covariates. However, it is often the case where we want to understand how certain covariates modulate dependence between certain endpoints. For instance, we may expect the covariate age or primary site to account for the apparent association between breast cancer survival and ovarian cancer survival, while other covariates do not play a role.

Thus, it is desirable to have a model where each pair-copula depends on a subset of covariates. This can be accomplished using a structured prior on the neural network weights, thus inducing a type of sparsity, or by replacing the neural network with a generalized additive model (GAM). With a GAM each covariate's effect is decomposed into smooth and interpretable function components. A drawback of using the GAM approach in place of a neural network, however, may be that substantially non-linear modulations of dependence across covariates are less well represented; nonetheless, this "interpretability vs. capacity bias" trade-off may be worthwhile in some applications such as clinical decision support. Currently, the implementation learns the vine structure from scratch in each epoch via maximum spanning trees on pairs of dependence measures and dynamically updates the tree with each training iteration stymies effective analysis of changes in the conditional dependence graph across training and patient subpopulations. One potential extension of these ideas is to adopt a fully Bayesian approach to the vine structure that would marginalize over its uncertainties during prediction, resulting in a posterior distribution over edges that would amount to a principled quantity measuring the strength of the dependence.

However, this raises more difficult questions of model fitting, since a Bayesian treatment would require computation of integrals over a discrete space of regular vines—a combinatorial optimization problem that is not trivial to solve. Efficient methods for sampling from the space of regular vines remain a challenge worth pursuing. While the advantage of the deterministic saddle-point approximations is computational efficiency, the drawback is that these formulations obfuscate the stochastic nature of the relevant epistemic uncertainty via the omission of its random component. In a full Bayesian treatment, the posterior variance across latent failure times would be captured only through sampling, while the deterministic implementation discussed in this chapter approximates this variation through the Hessian-based Gaussian approximation (Equation 18). The approximated uncertainty interval was well-calibrated according to the examples shown; however, more work is warranted to measure the interval coverage in relation to samples from the full posterior distribution produced via a Markov chain Monte Carlo simulator, especially in the case of model misspecification, which may lead to systematic under coverage.

Extensions to Complex Censoring Regimes and Domain-Specific Applications

The SP-DGCSM we have discussed here assumes, for simplicity, that observations are right-censored with censoring times distributed independently of event times. However, many real-life situations lead to left-censoring, interval-censoring, and possibly competing terminal events. These patterns are summarized by Pintilie (2006)^[35], and our next challenge is to extend the SP-DGCSM to deal with them through the rich structure provided by the saddle-point approximation. The constrained saddle-point equations can be solved iteratively, yielding a computationally feasible way to evaluate the Laplace approximation. When failure times are interval-censored, the variables for which this occurs have lower and upper bounds, i.e., $[w_j^{min}, w_j^{max}]$ and $[l_j, r_j]$, and

simple changes of variables yield to hyper-rectangular integration domains. This is not a problem for a truncated Gaussian copula, where one can simply integrate over the truncated multi-normal distribution. But the saddle-point equations must be solved subject to the box constraints imposed by the interval censoring. Fortunately, the saddle-point equations can be solved without the constraints, and the approximation is simply clipped to remain feasible (possibly in the saddle-point equations themselves). Wongkamthong and Akande (2023)^[43] employ variations on this clipping technique to handle imputation of ordinal data, and the same techniques may be effective here. The saddle point equations begin as in Eq 12 and then start clipping the approximation to the feasible box. When clipping leads to a boundary solution, the constrained saddle point will lie on the boundary of the box constraint, yielding a lower-dimensional integral over the unconstrained components only.

The class of problems one might have in mind when seeking the type of augmented saddle-point approximations considered in this paper would be situations arising in competing risks. In such problems only the minimum observed time of failure is known, together with its cause; the failure times associated with the other causes remain hidden. The likelihood in such cases must integrate over all the associated times for the different causes. Fortunately, a multivariate model for survival times provides a natural framework since one can view the dependent failure times in this situation. The interested reader should follow recently developed models based on vine copula approaches, which enable one to model the joint distribution of such dependent failure times without needing to assume independence of causes. The saddle-point approximation can then be applied to the latent times associated with the hidden causes. Such a model may well outperform models based on independent causes that dominate much of the literature (Pintilie, 2006)^[35]. The classic situation considered so far is that covariates are fixed throughout an individual's lifetime. The SP-DGCSM can be extended to incorporate time-varying covariates and models of longitudinal survival data. For instance, the hazards associated with time-dependent biomarkers can be modeled through a neural hazard net (Kleinbaum & Klein, 1996)^[21], or a piecewise constant covariate that is measured only at clinic visits can be modeled with a piecewise linear cumulative hazard, and the saddle point solver can then be used independently within each segment.

When multiple biomarkers are observed, one can model them as a multivariate stochastic process with copulas capturing contemporaneous dependencies. One popular strategy for modeling such multivariate time-dependent survival patterns is through dynamic vine copulas (Nagler *et al.*, 2022)^[30], where correlation parameters associated with each pair copula will evolve according to observed histories. Again, the saddle-point approximation could be employed at each time step to support differentiable learning of the resulting dynamic survival dependencies. The SP-DGCSM can be applied to numerous survival-analysis settings, including in oncology and critical-care medicine (such as was done for SEER and MIMIC-III), reliability engineering, finance, and demography. The reliability-engineering variant of the model estimates the joint failure times of an entire system's

components based on partial observations of system failures. In a finance context, the model specifies the joint default time of a portfolio of firms, a classic problem originating in the field of credit risk (Altman *et al.*, 1998)^[2]. In demography, versions of the model may be used in the study of the joint survival patterns of families and/or populations, accommodating genetic and environmental correlations. In any of these high-dimensional, consequential applications, the SP-DGCSM's scalable and deterministic saddle-point approximation ensures computational feasibility and prediction accuracy.

Conclusion

Lastly, we propose the Saddle-Point Deep Generative Copula Survival Model (SP-DGCSM), a novel approach that addresses the critical challenge in learning high-dimensional dependence structures, particularly for right-censored survival data. This principled procedure, based on SP-DGCSM, relies on a deterministic saddle-point approximation of the surrogate likelihood, instead of a standard stochastic integral approximation. The censored likelihood can be rewritten as a path integral (Cameron-Martin theorem) over latent variables uniformly distributed in the unit sphere with unit exponential distributions. We then identify the saddle point, yielding the dominant contribution through the steepest descent method. A second-order Taylor expansion at the saddle point leads to a closed-form expression of the Gaussian integral. Automatic differentiation is then applied to compute the resulting exact gradient, yielding superior stability for gradient-based optimization, leading to efficient learning even in high-dimensional spaces. Remarkably, the recursive saddle-point approximations serve to greatly mitigate the bivariate pair-copula problem by facilitating the computation of the latent integrals over one dimension with closed-form expressions. This architecture jointly learns both the neural marginal hazards and the vine correlation parameters in an end-to-end manner, with a preconditioned optimizer making use of the Hessians to further enhance convergence speed. The performance on synthetic and real-world examples (including a 500-dimensional stress test) typically exceeds the state-of-the-art in terms of concordance, calibration, and joint likelihood, while being suitable for uncertainty quantifications by virtue of the determinism in the inference—two-dimensional calibration curves are almost indistinguishable from the empirical estimates in the extreme quantile regions. The SP-DCGM provides an elegant solution to common survival analysis challenges, including handling right-censoring and discovering latent structures for predictive variables, all while avoiding ad hoc sampling from distributions. These models elegantly combine rich statistical capabilities for fitting arbitrary dependence structures with the computational tractability often desired in high-dimensional applications. Future work could focus on implementing the SP-DCGM with other common scenarios in survival analysis, such as interval-censoring, competing risks, and dynamic covariates. We also envision future research aimed at exploring interpretable neural-parameterized dependence structures.

References

1. Aas K, Czado C, Frigessi A, Bakken H. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*. 2009;44(2):182–198. doi:10.1016/j.insmatheco.2007.02.001
2. Altman EI, Caouette JB, Narayanan P. Credit-risk measurement and management: The ironic challenge in the next decade. *Financial Analysts Journal*. 1998;54(1):7–11. doi:10.2469/faj.v54.n1.2140
3. Anil R, Gupta V, Koren T, Regan K, Singer Y. Scalable second order optimization for deep learning. *arXiv*. 2020. doi:10.48550/arXiv.2002.09018
4. Apellániz PA, Parras J, Zazo S. Leveraging the variational Bayes autoencoder for survival analysis. *Scientific Reports*. 2024;14:24567. doi:10.1038/s41598-024-76047-z
5. Arrieta-Ibarra I, Gujral P, Tannen J, Tygert M, Xu C. Metrics of calibration for probabilistic predictions. *Journal of Machine Learning Research*. 2022;23(351):1–54. Available from: <https://www.jmlr.org/papers/v23/22-0658.html>
6. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. *Journal of the American Statistical Association*. 2017;112(518):859–877. doi:10.1080/01621459.2017.1285773
7. Blondel M, Berthet Q, Cuturi M, Frostig R, Hoyer S, Llinares-López F, *et al.* Efficient and modular implicit differentiation. *Advances in Neural Information Processing Systems*. 2022;35. doi:10.52202/068431-0378
8. Cam Le LM. Asymptotic methods in statistical decision theory. Springer; 2012. doi:10.1007/978-1-4612-4946-7
9. Chen RTQ, Rubanova Y, Bettencourt J, Duvenaud DK. Neural ordinary differential equations. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, editors. *Advances in neural information processing systems* 31. Curran Associates, Inc.; 2018. p. 6571–6583. Available from: <https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html>
10. Crawford SL. An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association*. 1994;89(425):259–267. doi:10.1080/01621459.1994.10476467
11. Dinh L, Sohl-Dickstein J, Bengio S. Density estimation using real NVP. *International Conference on Learning Representations*. 2017. Available from: <https://openreview.net/forum?id=HkpbnH9lx>
12. Fernandez C, Chen CS, Gaillard CP, Silva A. Experimental comparison of ensemble methods and time-to-event analysis models through integrated Brier score and concordance index. *arXiv*. 2024. Available from: <https://arxiv.org/abs/2403.07460>
13. Gerg ID, Monga V. Structural prior driven regularized deep learning for sonar image classification. *IEEE Transactions on Geoscience and Remote Sensing*. 2021. Available from: <https://arxiv.org/abs/2010.13317>
14. Glasserman P. Monte Carlo methods in financial engineering. Springer; 2003. doi:10.1007/978-0-387-21617-1

15. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, *et al.* Generative adversarial nets. *Advances in Neural Information Processing Systems*. 2014;27:2672–2680. Available from: <https://proceedings.neurips.cc/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html>
16. Janke T, Ghanmi M, Steinke F. Implicit generative copulas. *Advances in Neural Information Processing Systems*. 2021. Available from: <https://arxiv.org/abs/2109.14567>
17. Joe H, Cooke RM, Kurowicka D. Regular vines: Generation algorithm and number of equivalence classes. In: Kurowicka D, Joe H, editors. *Dependence modeling: Vine copula handbook*. World Scientific Publishing; 2011. p. 219–231. doi:10.1142/9789814299886_0010
18. Kalbfleisch JD, Prentice RL. *The statistical analysis of failure time data*. 2nd ed. Wiley; 2002. doi:10.1002/9781118032985
19. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv*. 2014. doi:10.48550/arXiv.1412.6980
20. Kingma DP, Welling M. Auto-encoding variational Bayes. *arXiv*. 2013. doi:10.48550/arXiv.1312.6114
21. Kleinbaum DG. *Survival analysis: A self-learning text*. Springer; 1996. doi:10.1007/978-1-4757-2555-1
22. Kreuzer A, Czado C. Bayesian inference for dynamic vine copulas in higher dimensions. *arXiv*. 2019. Available from: <https://arxiv.org/abs/1911.00702>
23. Krishnapuram B, Raykar VC, Steck H, Dehing-Oberije C, Lambin P. On ranking in survival analysis: Bounds on the concordance index. In: Platt JC, Koller D, Singer Y, Roweis ST, editors. *Advances in neural information processing systems 20*. Curran Associates, Inc.; 2007. p. 1209–1216. Available from: https://proceedings.neurips.cc/paper_files/paper/2007/file/33e8075e9970de0cfea955afd4644bb2-Paper.pdf
24. Lee C, Zame W, Yoon J, van der Schaar M. DeepHit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018;32(1). doi:10.1609/aaai.v32i1.11842
25. Li X-D, Li M-M. A novel nomogram to predict mortality in patients with stroke: A survival analysis based on the MIMIC-III clinical database. *BMC Medical Informatics and Decision Making*. 2022;22:92. doi:10.1186/s12911-022-01836-3
26. Ling CK, Fang F, Kolter JZ. Deep Archimedean copulas. *Advances in Neural Information Processing Systems*. 2020;33. Available from: <https://proceedings.neurips.cc/paper/2020/hash/10eb6500bd1e4a3704818012a1593cc3-Abstract.html>
27. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*. 1989;45:503–528. doi:10.1007/BF01589116
28. Magnus JR. *Matrix differential calculus with applications in statistics and econometrics*. 3rd ed. Wiley; 2019. doi:10.1002/9781119541219
29. Moolgavkar SH, Chang ET, Watson HN, Lau EC. An assessment of the Cox proportional hazards regression model for epidemiologic studies. *Risk Analysis*. 2018;38(4):777–794. doi:10.1111/risa.12865
30. Nagler T, Krüger D, Min A. Stationary vine copula models for multivariate time series. *Journal of Econometrics*. 2022;227(2):305–324. doi:10.1016/j.jeconom.2021.11.015
31. Nagpal C, Li X, Dubrawski A. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*. 2021;25(8):3163–3175. doi:10.1109/JBHI.2021.3052441
32. Nelsen RB. *An introduction to copulas*. 2nd ed. Springer; 2006. doi:10.1007/0-387-28678-0
33. Norcliffe A, Cebere B, Imrie F, Lió P, van der Schaar M. SurvivalGAN: Generating time-to-event data for survival analysis. *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*. 2023;206:10279–10304. doi:10.48550/arXiv.2302.12749
34. Othus M, Li Y. A Gaussian copula model for multivariate survival data. *Statistics in Biosciences*. 2010;2(2):154–179. doi:10.1007/s12561-010-9026-x
35. Pintilie M. *Competing risks: A practical perspective*. Wiley; 2006. doi:10.1002/9780470870709
36. Qi Y. *Extending expectation propagation for graphical models [dissertation]*. Cambridge (MA): Massachusetts Institute of Technology; 2005. Available from: <https://dspace.mit.edu/handle/1721.1/33069>
37. Rezende DJ, Mohamed S. Variational inference with normalizing flows. In: Bach F, Blei D, editors. *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. PMLR; 2015. p. 1530–1538. Available from: <https://proceedings.mlr.press/v37/rezende15.html>
38. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2009;71(2):319–392. doi:10.1111/j.1467-9868.2008.00700.x
39. Saikia R, Barman MP. A review on accelerated failure time models. *International Journal of Statistics and Systems*. 2017;12(2):311–322. Available from: http://www.ripublication.com/ijss17/ijssv12n2_15.pdf
40. Times S. SEER cancer statistics review, 1975–2008. National Cancer Institute; 2011. Available from: https://seer.cancer.gov/archive/csr/1975_2008/
41. Wiegand S, Kopper P, Sonabend R, Bischl B, Bender A. Deep learning for survival analysis: A review. *Artificial Intelligence Review*. 2024;57:65. doi:10.1007/s10462-023-10681-3
42. Wienke A. *Frailty models in survival analysis*. Chapman & Hall/CRC; 2010. doi:10.1201/9781420073911
43. Wongkamthong C, Akande O. A comparative study of imputation methods for multivariate ordinal data. *Journal of Survey Statistics and Methodology*. 2023;11(1):189–212. doi:10.1093/jssam/smab028
44. Wood SN. *Generalized additive models: An introduction with R*. 2nd ed. Chapman & Hall/CRC; 2017. doi:10.1201/9781315370279

45. Wu H, Su X, Singh VP, Feng K, Niu J. Agricultural drought prediction based on conditional distributions of vine copulas. *Water Resources Research*. 2021;57(8):e2021WR029562. doi:10.1029/2021WR029562
46. Zhong L, Yang F, Sun S, Wang L, Yu H, Nie X, *et al.* Predicting lung cancer survival prognosis based on the conditional survival Bayesian network. *BMC Medical Research Methodology*. 2024;24:16. doi:10.1186/s12874-023-02043-y

How to Cite This

Nangolo PN. Phenotype-structured modelling of tumour plasticity and drug resistance under adaptive therapy. *Int J Appl Math Numer Res*. 2026;2(5):12-21.

Creative Commons (CC) License

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution NonCommercialShareAlike 4.0 International (CC BYNC-SA 4.0) License, which allows others to remix, tweak, and build upon the work noncommercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.